Universitat de Barcelona

Ecology and evolution of microbial nitrifiers

Ecología y evolución de los microorganismos nitrificantes

Antonio Fernàndez Guerra

CSIC
Consejo Superior de Investigaciones Científicas

Ecology and evolution of microbial nitrifiers

A. Fernàndez Guerra

Ecology and evolution of microbial nitrifiers

Antonio Fernàndez Guerra

PhD Thesis 2012
University of Barcelona

Universidad de Barcelona
Facultad de Biología

Ecology and evolution of microbial nitrifiers

Ecología y evolución de los microorganismos nitrificantes

Antonio Fernàndez Guerra

Tesis Doctoral

Universidad de Barcelona

Facultad de Biología

Programa de Doctorado de Genética

A genetic approach to the Ecology and Evolution of microbial nitrifiers

Memoria presentada por Antonio Fernàndez Guerra
para optar al Grado de Doctor por la
Universidad de Barcelona

Antonio Fernàndez Guerra
Centro de Estudios Avanzados de Blanes (CEAB)
Consejo Superior de Investigaciones Científicas (CSIC)

Barcelona, Diciembre de 2012

El director de la tesis
Dr. Emilio Ortega Casamayor
Inv. CEAB–CSIC

El tutor de la tesis
Dr. Julio Rozas Liras
Prof. UB

Para los que siempre habéis estado ahí.

Agradecimientos

We are drowning in information but starving for Knowledge

John Naisbitt

Al hacer memoria y recordar todos los momentos pasados... todos los nuevos amigos... todas las experiencias que he acumulado a lo largo del camino... uno se emociona y a veces hasta le saca una sonrisa picarona... aunque sea entre los *suaves* tonos azul y amarillo de un Boeing 737-800 de Ryanair cruzando el Mar del Norte hacinado en un minúsculo espacio y aburrido de que te vendan de todo y más... pero por mi fortuna... el tiempo durante los que he estado gestando esta tesis ha sido totalmente diferente... ha sido mucho mejor que hubiese podido esperar...

Como siempre me gusta decir, esto para mi no es nada más que un juego, la ciencia es mi hobby y por suerte lo he podido convertir en un oficio... me dan euros *euros* por disfrutar... que más puedo pedir... bueno... sí... siempre se puede pedir más... si a todo esto le añadimos playa y buen tiempo... y tener como compañeros de despacho a unos colegas... ¡ya es la bomba! Muy bien... Quién lea esto, va a pensar... ...vaya, todo muy bonito... pero ¿y la ciencia qué? Pues muy bien, gracias. Salvando las distancias, nuestro pequeño despacho es como el *Googleplex* del CEAB... entre material de escalada, piragüismo y mucho trasto por identificar se esconde un laboratorio de ideas donde unos individuos con medios limitados han sido capaces de hacer una ciencia novedosa y de vanguardia. Cómo dice la canción *The time of my life...*

Como soy bien agradecido ha llegado el turno de dar las gracias a todos aquellos que de una forma u otra han sido participes de una pequeña pero muy intensa parte de mi vida.

Como no, el primero en agradecer es Emilio, ya que sin tu confianza nunca hubiese estado escribiendo estas líneas. Gracias por dejar de lado todo el *establishment* académico basado en un simple número, seguir tu instinto y apostar por un estudiante que de buenas a primeras no destacaba mucho...

gracias por darme la libertad de hacer lo que he querido durante estos años para satisfacer mi curiosidad aunque no tuviese nada que ver con el tema de la tesis. Gracias por guiarme y aconsejarme durante los inicios de mi carrera científica. Y gracias por darme la oportunidad de conocer a la buena gente que he ido encontrando a lo largo del camino. Y muchas gracias por el último esfuerzo que has dedicado para ayudarme a acabar esta tesis cuando ya estaba lejos del CEAB empezando mi nueva etapa científica en Bremen.

Los siguientes a agradecer son un par de individuos con los cuales he compartido estos años montones de momentos buenos y algún que otro no tan bueno (pero por suerte no muchos). La verdad que estos chavales no tienen precio y nos lo hemos pasado en grande. Tantas cosas que contar... bueno mejor no todas que sino la liamos... Lo mejor de la tesis... ellos y con diferencia. Ahora como en las grandes ocasiones ahí van los agradecimientos uno a uno. Gracias Jose Christopher, digo francés de... digo Jean Christophe Auguet por ayudarme en todo lo relacionado a los *bichitos* estos que oxidan el amonio... pero mucho más importante que eso, por ser un buen amigo, por las salidas en *kayak* mareando a tu cuñado, por todas las cervezas y tapas que hemos compartido, por esa jocosidad que tienes escondida detrás tu *savoir faire* francés... En la agenda tengo apuntada la visita a Pau que te debo.

Bueno... bueno... ahora es el turno del Sr. Albert Barberán Torrents, el Dr. Barberán... vaya... ahora ya le dan el título a cualquiera... no... *si el que más tonto parece es el más listo*... o algo parecido dijo alguien con muchas luces en uno de los desayunos en la terraza del CEAB. La verdad que tenía razón, en lo de listo. Creo que Emilio nunca hubiese pensado encontrar una pareja que se complementarán tan bien... en ciencia... por mucho que lo intentes mi codo me mantendrá a salvo de tus intentos de *spooning*. La verdad que nos lo hemos pasado muy bien jugando a hacer ciencia. Atrás quedan muchos recuerdos y risas... muchas *Patrullas Águila* para entrar al *Lile Eule*... *Albert, ha sigut un plaer compartir tots aquests anys tant a nivell científic com a nivell personal. No t'he de desitjar sort en la nova etapa que comences a Colorado, ets un màquina nano!*

En el grupo han ido y venido muchos compañeros, pero quien tiene un lugar especial es Juan. Gracias a él empecé en serio a la bioinformática cuando aún era un estudiante de biología, ayudándolo con los *dnawalks* y la fractalidad en los genomas. Nunca compartimos despacho... bueno... de hecho... no teníamos... pero eso no evitaba que durante los trayectos en coche cuando íbamos a ver Emilio los Jueves tuviésemos charlas interesantes y preparásemos los experimentos que hacer. Mucha suerte en Viena y te deseo lo mejor para ti

y tu familia. ¡Seguro que Artur ja se ha echado novia!

Luego estaba esa chica de un pequeño pueblo de León, Carmen, que ponía la cordura en el despacho... santa paciencia... con los tres piezas que te tocaron. Hemos tenido nuestros más y nuestros menos... incluso rumores... Te deseo suerte con tus andaduras por tierras catalanas... y ¿qué va a hacer ahora el niño por las américas sin ti?

Estos últimos años hemos tenido un par de nuevas incorporaciones... Claudia, una loca colombiana con añoranza de su tierra que ha conquistado a un francés... de los mejores inicios que recuerdo... impresionante la primera cena en mi casa... dejaste el listón muy alto.

Luego vino Maria Vila Costa, *una bona amiga de bona pasta, espero que tinguis molta sort amb la nova etapa que comences com a mare, i molta sort per en Soori... a veure si podem anar finalment a Montblanc a matar el Drac.*

Tomàs, la nueva incorporación del despacho, un *freak* del *running*... *Encara esperem a que ens invitis als dinars que feia la teva mare... Molta sort amb la metagenòmica i la bioinformàtica, és més fàcil del que sembla*.

Y el último en llegar fue Xevi. *El noi d'Anglès, que un dia em va donar classe a la uni... un bonàs, una gran persona i un gran treballador. Et desitjo a tu i a la Teresa que tingueu molta sort!*

En el CEAB he tenido la suerte de conocer gente muy maja, incluso he hecho algún que otro amigo. Sin lugar a dudas, *l'hereu de can Oller. Quin parell de losers ens hem anat a trobar... Miquel, merci per aquestos anys a Blanes, ens ho hem passat molt bé entre calçotades, rallies, Bremen, gin tonics i flipar amb la fauna varia del CEAB*.

Luego hay esa pareja de madrileños con los que iniciamos unas muy buenas cenas los Jueves en mi terraza. ¡Edu y Clara que tengáis mucha suerte! Edu a mi ya no me vuelves a engañar con hacer presas... Steffi, la alemana escaladora enamorada de mi tierra, que me ayudó a aprender un poco de alemán antes de irme por primera vez a Bremen. Y suerte que volvió Fede, un soplo de aire fresco, una mente privilegiada... *Em moro de ganes de poder començar d'una vegada els projectes que tenim entre mans.*

También hay que agradecer a los que han financiado mi tesis. *Gràcies Jordi per confiar amb l'Emili, apostar per mi i finançar-me durant un bon temps*. Durante estos años he ido de proyecto en proyecto... Agradezco parte de mi financiación y formación a:

- CRENYC CGL2006-12058-CO2-02/BOS (MEC)
- PIRENA CGL2009-13318-C02-01/BOS (MICINN)
- DARKNESS CGL2012-32747/BOS (MINECO)

- Graccie Ref. CSD2007-00067 (Consolider-Ingenio 2010)
- European Cooperation in Science and Technology-COST. Action number: ES1103 CISME Ref OC-2010-2-8674
- EU-FP6-NETWORK OF EXCELLENCE MARINE GENOMICS GOCE-CT-2003-505403
- Acciones Integradas Alemania-España HD2008-0006 (MICINN)
- Ecogenomica comparativa microbiana EO-CSIC (CESCA)
- BCV-2010-3-0003: Diversity and evolution of marine microbial communities investigated using high throughput DNA sequencing technologies (metagenomics, genomics & phylogenomics)

Gracias a esta financiación he podido viajar y realizar estancias en Francia y Alemania. La verdad que a parte de ser productivas, he conocido a gente increíble. I had a great summer at *l'Observatoire Océanologique de Banyuls*. I met such a nice guys... My warmest thoughts to Mr Cristal, 23 pack and Matan. Hope we will meet some day in the future to enjoy more Mojitos and Cuban Cigars... And many thanks to Anne Marie, the *Sea Urchin* experience was great, I owe you a lot.

And a few years ago, I had the chance to visit the *Max-Planck-Institut für Marine Mikrobiologie in Bremen* (I never thought I'd end up living there). Finally I am in a research group full of geeks... During those years I met such a nice people there, such a nice friends. I don't have enough words to express my gratitude to Pelin, Sandra, Julia, Ivo, Daphne, Gerd, Ana, Sven, Pablo, Paola... Well when this will be finished, beers are on me! But there is a special place on my little heart for three of the best human beings I know... Melissa, Petra and Renzo I am so happy that our lifes crossed at some point...

Pero por suerte, hay más mundo fuera del CEAB y durante estos años de doctorado he coincidido con viejos amigos que me han hecho este trayecto mucho más agradable. Sin ir más lejos en Blanes un día andando por la calle me encontré con una vieja amiga del tiempo de la universidad que hacía años que no veía, era de Blanes y casi vecina. Gracias a ella y a sus amigos conocí los entresijos del pueblo. *Gràcies Maria, David, Blanca i Miquel per tots els sopars, caminades i converses que hem tingut durant aquests anys. I gràcies pel dia de la despedida, va ser molt bonic!*.

Por supuesto, uno siempre arrastra esos amigos de la universidad, que siempre estan ahí para lo que sea... o para dar la lata... Pasarán los años y

aún nos seguiremos encontrando y echando risas y haciendo el tonto... *El Dr. Nadales, qui ho havia de dir que en fariem res de bo de tu... i mirate'l amb una empresa i tot ara... Molta sort nano amb les teves plantes màgiques per terres andaluses. I que dir del Dr. Caliz i la Marina, la meva familia adoptiva de Girona... gràcies per tots aquests anys que sempre heu estat allà i molta sort amb l'aventura de ser papes*.

Luego están mis amigos de toda la vida de Alcover, que no son pocos... ahí van unos cuantos nombres, y seguro que me olvido de alguno, pero no me lo tengáis en cuenta... Iker, Cisco, Manel, Perru, Gallofas, Koixi, Ivan, Batet, Bagué, Ciuró, Edgar, Antón, Eva, Cristobal, Masqué, Antonio, Natalia, Abraham, Eva, Secu, Chichillo, Maria, Piti, Ramón... que tengáis mucha suerte en todo lo hagáis.

I was a lucky guy and part of my thesis has been written in *Tuulispää*, the farm where the dreams come true, in the middle of nowhere, surrounded by nature... and with *Kultaseni* taking care of me. I am really lucky to meet someone so strong and brave like you Piia... you gave me enough strength to finish that. *Haleja ja suukkoja Piia!*

En último lugar, pero sin duda los más importantes, mi **familia**, en especial a **mis padres** y **hermano**. Sin su paciencia ni su ayuda nunca hubiése sido capaz de poder ser lo que soy ahora. Sólo os voy a dedicar unas pocas palabras, pero sabed que el número de palabras que os dedico son inversamente proporcionales al amor que os profeso. Gracias de todo corazón.

En algún lugar sobre las nubes entre Suecia y
Finlandia. Noviembre de 2012

Contents

Resumen

Ecology and Evolution of Microbial Nitrifiers

Resumen

1 Resumen

1.1 Introducción general

El nitrógeno (N) es uno de los elementos más importantes para la vida en la Tierra, muchos de los compuestos esenciales en los procesos requeridos para la vida incluyen N como componente crítico, nucleótidos y aminoacidos son un buen ejemplo. El 80 % de la atmósfera esta formado por nitrógeno gas (N_2), pero a causa del triple enlace, esta reserva de N no puede ser utilizada por la mayoría de organismos en la Tierra. Este triple enlace se tiene que romper, pero es una reacción que requiere un alto nivel de energía. El proceso por el cual se obtiene nitrógeno reactivo (Nr) al romper el triple enlace se llama fijación del nitrógeno. Aunque la fijación del nitrógeno puede ser llevada a cabo por procesos naturales como los relámpagos o en los sistemas de altas temperaturas, como los que podemos encontrar en las chimeneas submarinas (Canfield et al., 2006), la cantidad de Nr no es suficiente para abastecer todas las formas de vida que se han ido desarrollando a lo largo de la historia de la Tierra. La gran parte del Nr proviene de la fijación mediada por microorganismos. Estos microorganismos han desarrollado una maquinaria metabólica especial para poder romper el triple enlace del N_2 y generar formas reducidas de N biológicamente activas que posteriormente podrán ser asimiladas por otros organismos.

Durante millones de años el Nr que provenía de la fijación biológica era uno de los factores para el desarrollo limitantes en la biosfera. Esta limitación era debida a la competición que había entre las diferentes formas de vida, y ha acabado desarrollando la biodiversidad actual y las relaciones entre los organismos. Este escenario cambió totalmente cuando los químicos descubrieron el papel esencial del nitrógeno para la bioquímica de la vida, y otros científi-

cos lo identificaron como un nutriente esencial para plantas y animales (Smil, 2004). En paralelo, hubo un creciente interés sobre la demanda de comida por la población mundial, la tasa de crecimiento de la cual sobrepasaba las tasas de producción de alimentos. Pero este escenario cambió después de que se revelara la fijación biológica de N. Una vez se sentaron las bases de los procesos naturales para la formación de Nr, se pudo desarrollar el proceso químico para convertir el N_2 a NH_3. En 1913, se inventó el proceso de Haber-Bosch, y por primera vez en la Tierra, existió la posibilidad de disponer de suministros ilimitados de Nr; éste fue el inicio de los fertilizantes artificiales. Existe una correlación entre el crecimiento de la población mundial y el descubrimiento del proceso de Haber-Bosch. Pero el proceso de Haber-Bosch no fue la única fuente antropogénica de Nr. Desde la revolución industrial, la energía producida por el uso de combustibles fósiles inyectaba Nr directamente en la atmósfera.

La actividad humana ha desestabilizado totalmente el ciclo del N y por primera vez en la historia de la Tierra, el Nr se encontraba en exceso. La tasa de generación de Nr excedía de largo la conversión a N_2 debido a los procesos de desnitrificación y el Nr se empezó a acumular en los ecosistemas. La acumulación de Nr en la naturaleza es un asunto que puede acarrear serias consecuencias para los humanos y los ecosistemas; los efectos de esta acumulación pueden ir desde la eutrofización y acidificación de los ecosistemas acuáticos y terrestres a la pérdida de ozono en la estratosfera, entre otros muchos problemas (Galloway & Cowling, 2002). De todas formas la principal preocupación del exceso de Nr, es el conocido *efecto cascada*. De Gruber & Galloway (2008):

> Por ejemplo, una molécula emitida de oxido nitroso puede causar en primera instancia *smog* fotoquímico y luego de ser oxidada en la atmósfera a ácido nítrico y depositada en los suelos, puede conducir a la acidificación y eutrofización de los ecosistemas

Pero la complejidad del problema aumenta cuando intentamos entender como los cambios en el ciclo del N afectan a los demás ciclos biogeoquímicos, y en particular al ciclo del carbono (C), ya que ambos están directamente relacionados. Por un lado, los factores antropogénicos están relacionados con el constante incremento de la disponibilidad del Nr y de la concentración de CO_2 atmosférico, uno de los causantes del calentamiento global. Existe una relación directa entre la fertilización de los campos y el CO_2 consumido por la agricultura intensiva, donde los niveles de CO_2 atmosférico disminuye mientras la biomasa vegetal aumenta (Schimel et al., 2001). Por otro lado, el ciclo del nitrógeno y del Carbono están íntimamente relacionados a causa de los procesos relacionados a la vida. Nitrógeno, carbono, fósforo y otros elementos son utilizados para ensamblar los bloques esenciales de la vida, que acabarán formando parte de los diferentes organismos que habitan el planeta. La unión

entre los diferentes elementos ocurren a estequiometrías específicas que determinan los vínculos entre los diferentes ciclos biogeoquímicos (Sterner & Elser, 2002).

Por ejemplo, la proporción de C/N es diferente entre organismos fotosintéticos marinos y terrestres, mientras que en los marinos existen pequeñas fluctuaciones en la proporción, en los terrestres las fluctuaciones son mucho mayor.

La cantidad de Nr en la Tierra está controlado por los procesos biológicos de fijación y desnitrificación, pero las alteraciones en estos procesos pueden afectar al ciclo global del Carbono y al clima, mientras tanto la proporción de C/N en los autótrofos puede continuar inmutable.

Aún no hay un acuerdo sobre como la fijación biológica y la desnitrificación están equilibradas en los océanos (Gruber & Galloway, 2008; Codispoti, 2007) pero lo que está claro es, (i) el ciclo del nitrógeno marino es muy dinámico, la tasa de renovación es menor de 3000 años (Gruber & Galloway, 2008); y (ii) el ciclo del fósforo marino es esencial para estabilizar el ciclo del N marino (Deutsch et al., 2007). En cambio, los sistemas terrestres no están tan bien estudiados, pero existe un gran transporte lateral de los suelos, donde las fuentes de Nr, exceden a la desnitrificación, a los sistemas de agua dulce, donde la desnitrificación prevalece. De hecho, en los sistemas terrestres, casi la mitad de la desnitrificación sucede en los sistemas de agua dulce (Seitzinger et al., 2006).

Los procesos de fijación y desnitrificación están principalmente mediados por grupos específicos de microorganismos. Con todas las perturbaciones introducidas en el ciclo del N por las recientes actividades humanes y los cambios inducidos a escala global, el estudio y la comprensión de las comunidades microbianas directamente relacionadas con el ciclo del N se ha convertido en una de las principales áreas de estudio de la ecología microbiana. Actualmente entender y expandir el conocimiento de los procesos implicados en la nitrificación-desnitrificación son cruciales para mantener el equilibrio entre los diferentes ciclos biogeoquímicos. Una de las etapas limitantes de la denitrificación natural es el continuo flujo de nitrato a través de la nitrificación. En la reacción de nitrificación, el NH_3 que proviene de la fijación artificial o natural de nitrógeno y de fuentes orgánicas o atmosféricas, son convertidas en nitrato a través de nitrito. Este nitrato podrá ser utilizada por los desnitrificadores o asimilados por los diferentes organismos. La nitrificación se realiza en dos etapas por dos grupos de organismos autótrofos fisiologicamente distintos: los oxidantes del amonio que oxidan el NH_3 a NO_2^-, y los oxidantes del nitrito, que oxidan el NO_2^- a NO_3^-. La oxidación del amonio es la etapa limitante de la nitrificación y es un proceso biogeoquímico de importancia global en ecosistemas naturales y artificiales alrededor del planeta que tienen un

sustancial impacto ambiental en las emisiones de gas con efecto invernadero (básicamente óxido nitroso N_2O, y oxido de nitrógeno NOx)

En suelos, la oxidación del amonio puede acarrear cuantiosas perdidas de nitrógeno debido a la desnitrificación o a la lixiviación de nitrato; las concentraciones de amonio en suelos han aumentado durante los últimos años a causa de los hábitos de uso del suelo y del incremento de las concentraciones atmosféricas de amonio (Rockström et al., 2009), estos procesos pueden llegar a modificar la ecología microbiana de los procesos de nitrificación (Verhamme et al., 2011).

En los ambientes marinos, la nitrificación corresponde a cerca de la mitad del nitrato consumido por el fitoplancton a escala global (Yool et al., 2007) y es el responsable de las reservas de nitrito en los océanos (Karl, 2007), la reserva más grande de nitrógeno reactivo de la biosfera (Beman et al., 2010; Leininger et al., 2006). También en los ambientes marinos, la oxidación del amonio es un componente importante de la mineralización del nitrógeno de origen orgánico y para la eliminación de las entradas antropogénicas de nitrógeno en las zonas costeras. En ecosistemas artificiales como los bioreactores o depuradoras, la oxidación del amonio es uno de los procesos clave para la eliminación del nitrógeno (Van Loosdrecht & Jetten, 1998; Mussmann et al., 2011). La nitrificación también puede ayaudar a eliminar el nitrógeno en exceso de los lagos y prevenir su eutrofización (Hagopian & Riley, 1998).

Las oxidadoras del amonio juegan un papel importante en la conexión del nitrógeno fijado por medios biológicos y las perdidad anaeróbicas, pudiendo ser detectados en una gran variedad de ambientes aquaticos y terrestres (Nicol & Schleper, 2006; Auguet et al., 2010). Tanto las oxidadoras del amonio bacterianas (AOB) como las arqueas (AOA) han colonizado ambientes similares en el planeta, pero con diferentes grados de éxito en términos de abundancia, actividad y distribución (Verhamme et al., 2011; Prosser & Nicol, 2008; Martens-Habbena et al., 2009; Auguet et al., 2011; Sauder et al., 2011).

La ecología molecular de la oxidación del amonio ha sido extensamente explorada gracias a estudios de la subunidad A (o α de la aminomonooxigenasa (Auguet et al., 2011; Rotthauwe et al., 1997; Agogué et al., 2008; Francis et al., 2005, 2007). Tanto los AOA como los AOB tienen el gen que codifica para la *amoA*; aunque, el gen ha evolucionado de forma distinta en cada uno de los dominios (Treusch et al., 2005).

La ammonio monooxigenasa (AMO) es la enzima clave para la conversión del amonio a la hidroxilamina, la cual será convertida en nitrito. La AMO está compuesta por tres subunidades,*amoA*, *amoB* y *amoC* y está evolutivamente relacionada a la metano monooxigenasa (pMMO) de las bacterias oxidadoras de metano (Holmes et al., 1995).

Oxidación del amonio en bacterias

En un principio se pensaba que la oxidación del amonio por parte de los microorganismos estaba restringida a un grupo de unas cuantas bacterias el filo de las *Proteobacterias* (γ and β), las cuales bajo las condiciones del laboratorio mostraban afinidad por concentraciones más altas de amonio que las encontradas en la naturaleza (Bollmann et al., 2002).

Durante muchos años los estudios ambientales moleculares de los AOB estaban centrados en el 16S rRNA de β asumiendo que los γ no estaban presentes, ya que estos AOB sólo habían sido detectados como aislados de ambientes marinos. Esta aproximación estaba totalmente sesgada ya que se basa en métodos dependientes de cultivos, pero en realidad la mayoría de microorganismos ambientales no crecían en cultivo puro. En el año 1993 McTavish et al. (1993) publicó la primera secuencia génica del *amoA* de un cultivo de *Nitrosomonas europaea*. A partir de entonces, la subunidad α del AMO fue utilizada como marcador y el número de las secuencias en las bases de datos asociadas a este gen aumentaron rapidamente.

El uso del gene del *amoA* en lugar del rRNA conllevó varias ventajas. Primero, los genes del *amoA* de las β- y γ-AOBs podían ser amplificados al mismo tiempo (Holmes et al., 1995; Mendum et al., 1999; Nold et al., 2000; Sinigalliano et al., 1995) o por separado (Rotthauwe et al., 1997; Stephen et al., 1999). Segundo, el *amoA* comparte un ancestro común con el gen de la subunidad α de la metano monooxigenasa (*pmoA*) (Holmes et al., 1995; Nold et al., 2000).

Las AOB se pueden encontrar normalmente en ambientes ricos en nitrógeno como bioreactores, biofilms y biofiltros. Durante muchos años los ecólogos microbianos estaban asombrados por la capacidad nitrificadora de determinados ecosistemas donde aparentement no se detectaban las AOB, sobretodo en condiciones oligotróficas (Olson, 1981).

Oxidación del amonio en arqueas

En los años 1990, como parte del censo de la diversidad bacteriana en uno de los reactores nitrificantes para tratar el agua en un acuario de agua salada en el the Shedd Aquarium (Chicago, Illinois), se formulo una hipótesis sobre la posible existencia de un nuevo grupo de oxidantes del amonio (Fuhrman et al., 1992) ya que la amplificación por PCR dirigidas a los grupos conocidos de bacterias oxidantes del amonio fallaron en detectarlas. Unos años después, se realizaron observaciones similares durante el estudio del procesado del nitrógeno de los sedimentos del Plum Island Sound (Massachusetts) (Könneke et al., 2005). Otra vez los cebadores específicos para bacteria fallaron en amplificar las secuencias esperadas. De todas formas, esta vez el estudio se amplió a

las secuencias del gen del 16S rRNA del clado I de las *Crenarchaeota*. Este clado era abundante en los enriquecimientos de los sedimentos de estuario y en los sistemas de filtración en el Shedd Aquarium (Könneke et al., 2005).

Estos resultados sugirieron que podía existir una relación entre el abundante Grupo I de las *Crenarchaeota* y la oxidación del amonio. El siguiente paso era obtener cultivos enriquecidos utilizando el material del acuario como inoculante, y añadir al medio de cultivo una concentración menor de amonio que la utilizada en los cultivos de bacterias oxidantes del amonio. Finalmente se pudo obtener un cultivo enriquecido de una población de arqueas afiliada al Grupo I de las *Crenarchaeota* y se pudo describir por primera vez un cultivo puro de una AOA y la cepa *Nitrosopumilus maritimus* strain SCM1 pudo ser descrita.

La combinación de métodos independientes de cultivo para asociar la oxidación del amonio, al Grupo I de *Crenarcheota* y el posterior aislamiento en cultivo, impulsó el cultivo de otras AOA de diferentes ambientes como *Nitrososphaera viennensis* aislada del suelo (Tourna et al., 2011) y *Nitrosotalea devanaterra* aislada de un suelo con bajo pH (Lehtovirta-Morley et al., 2011). A partir del cultivo de *Nitrosopumilus maritimus* y las otras AOA, la diversidad filogenética de las arqueas oxidantes del amonio ha sido revisada varias veces mientras nuevos organismos han sido descritos (Pester et al., 2012). Actualmente, las filogenias son construidas utilizando un gran número de secuencias del gen del 16S rRNA o de la subunidad α del *amoA* revelando nuevas relaciones filogenéticas como el clado *Thaumarchaeota*, una nueva división dentro de Archaea donde se encuentran todas las AOA descritas hasta la fecha (Brochier-Armanet et al., 2008; Pester et al., 2011; Spang et al., 2010). Una característica común de las *Thaumarchaeota* cultivadas es la de crecer oxidando el amonio. Las *Thaumarchaeota* al ser un grupo muy divergente dentro de las arqueas, están generando nuevas preguntas sobre el origen de las arqueas y sobre la oxidación del amonio, ya que los clados más ancestrales de Archaea están dominados por fisiologías anaeróbicas.

La distribución ambiental de las AOA es extremadamente diversa sobrepasando a la observada en las bacterias. La abundancia y los patrones de diversidad han sido obtenidos básicamente a partir de la secuenciación y la cuantificación del gen que codifica para el presunto gen del *amoA* (Francis et al., 2005; Beman et al., 2008). La subunidad A de la amonia monooxigenasa normalmente se encuentra bien correlacionada con la abundancia de crenarcheol (Damsté et al., 2002; Könneke et al., 2005; Leininger et al., 2006). Los resultados de Lehtovirta et al. (2009); Ochsenreiter et al. (2003) y Leininger et al. (2006) sugieren que las AOA parecen ser el clado dominante en suelos, abarcando hasta el 1-5 % de los procariotas y siendo el grupo marino dominante representando el 20-40 % del bacterioplancton marino (Karner et al., 2001; Church et al., 2003).

Las AOA también parecen ser la población más abundante de oxidantes de amonio en hábitats geotermales (de la Torre et al., 2008; Zhang et al., 2008; Reigstad et al., 2008). Las AOA cultivadas pueden llegar a crecer a temperaturas hasta 74ºC (*Nitrosocaldus yellowstonii*). El amplio abanico de hábitats en el que encontramos las AOA es reflejado en la tremenda diversidad filogenética definida por los diferentes tipos definidos por el gen del 16S rRNA reconocidos por tener afiliado oxidantes del amonio y por la correspondiente diversidad de las secuencias de *amoA* de arqueas.

Diferencias entre las AOB y las AOA: dos estrategias diferentes para la oxidación del amonio

Martens-Habbena & Stahl (2011) compararón la estequiometría de la oxidación del amonio entre *N. maritimus* y las AOB utilizando microespirmetría para medir el consumo de oxígeno y de amonio relativo al nitrito, y concluyó que la estequiometría global en AOA es indistinguible de la de las AOB:

$$1NH_3 + 1{,}5O_2 \Rightarrow 1NO_2^- + H_2O + H^+$$

En Bacteria, la amonio monooxigenasa es la enzima clave responsable para la conversión del amonio a hidroxilamina, la cual es convertida a nitrito por la hidroxilamina oxidoreductasa; pero el análisis genómico del *Thaumarchaeota* marinio Candidatus *Nitrosopumilus maritimus* SCM1, ha revelado la existencia de un sistema para la oxidación del amonio diferente de la utilizada por las AOB. En consecuencia, dos hipótesis fueron propuestas para explicar la falta del homólogo para el complejo de la hidroxilamina oxidasa y la capacidad para la síntesis de los citocromos tipo-c. Una hipótesis asume que el proceso puede estar conducido por una de las oxidasas multicobre periplásmicas; y la otra sugiere el uso del nytroxyl como intermediario en lugar de la hidroxilamina (Walker et al., 2010).

Además, sólo dos pequeñas proteínas similares a la plastocianina están compartidas por todas las AOA como resultado de análisis de genómica comparada (Stahl & de la Torre, 2012). En la hipótesis del nytroxyl, estas proteínas que contienen cobre podrían participar en la transferencia de electrones desde el nytroxyl a la cadena de transporte electrónica ligada a membrana. Aunque la vía de las arqueas para la oxidación del amonio no se ha podido resolver por genómica comparada, estudios recientes utilizando microelectrodos sensibles al óxido nítrico (NO) indican que el NO podría estar implicado en el proceso bioquímico (Stahl & de la Torre, 2012). Cantidades apreciables de NO son producidas durante la oxidación del amonio. Esto ha llevado a hipotetizar

que el NO podría ser un intermediario o funcionar como lanzadera redox, por ejemplo, entregando electrones a la AMO.

A diferencia de lo que sucede en las AOA, en las AOB, toman los electrones requeridos por la monooxigenasa del acervo de quinonas. Si en las AOA la formación de nytroxyl como primer producto de la oxidación del amonio o la utilización del NO con lanzadera redox para la formación de hidroxilamina, eliminaría el préstamo directo del acervo de quinonas, ya sea, obviando la necesidad de un agente reductor a través de la formación de nytroxyl or por el préstamo de electrones de un donador con un potencial electrónico menor en la reducción de nitrito a NO.

Como la formación de hidroxilamina como producto intermediario de la AMO aún no ha sido demostrada, la vía de las arqueas para la oxidación del amonio aún se considera no resuelta.

Ecología filogenética de los nitrificantes microbianos

Las diferencias ecológicas y fisiológicas y las similitudes entre las AOA y las AOB son el escenario perfecto para aplicar métodos de filogenia comparativa para analizar la estructura de comunidades y los patrones de diversificación utilizando la secuencia génica el *amoA*.

Desde el desarrollo del contraste filogenético independiente (PIC) (Felsenstein, 1985; Ackerly, 2009) se ha sucedido una revolución el campo de la ecología de comunidades (Webb et al., 2002). Ecología y evolución están íntimamente asociadas y algunas metodologías han sido desarrolladas para analizar las constricciones observadas en la naturaleza. Actualmente, los ecólogos de comunidades pueden evaluar donde la mayor parte de la diversidad biológica se acumula (Faith, 1992) y como esta diversidad está estructurada (Webb, 2000; Helmus et al., 2007); o como la β-diversidad filogenética está distribuida a lo largo de los gradientes ambientales (Lozupone & Knight, 2005; Bryant et al., 2008; Ives & Helmus, 2010).

La combinación de la información filogenética con las medidas tradicionales de la ecología como la la riqueza específica, han derivado en medidas mucho más completas como la Diversidad Filogenética (PD) (Faith, 1992). La PD mide la diversidad de los linajes como la suma de la longitud de las ramas de los miembros en una comunidad, añadiendo a la riqueza específica la historia evolutiva. Un valor mayor de PD indica que las especies en una comunidad están menos relacionadas en términos filogenéticos. Métricas como la distancia media por parejas (MPD), que miden la media de la distancia filogenética entre todos los miembros de cada comunidad, o la distancia media al taxón más cercano (MNTD) que calcula la distancia media que separa cada miembro de su pariente más cercano (Webb, 2000; Webb et al., 2002)

son útiles para examinar los motores ecológicos y evolutivos del ensamblaje de la comunidad. Por otro lado, el índice de variabilidad de especies (PSV) cuantifica como la relación filogenética decrece la diferencia de una hipotética característica neutra compartida por todos los miembros de la comunidad. Cuando el valor es 1, todas las especies no están relacionadas, mientras que cuanto más se aproxima a 0 las especies se encuentran más relacionadas (Helmus et al., 2007). Medidas de β-diversidad o similitud de comunidades son definidas como la fracción de la longitud de la ramas compartida por dos comunidades (Lozupone & Knight, 2005; Bryant et al., 2008). La β-diversidad taxonómica y la β-diversidad filogenética será la misma si en dos comunidades todas las especies están igualmente relacionadas, por ejemplo, en un filogenia estrellada.

Las filogenias derivadas de datos moleculares proporcionan un registro indirecto de los episodios de especiación que llevaron a cabo las especies extintas, reflejando el tempo y el modo de los procesos microevolutivos y macroevolutivos relacionados a la diversificación (Mooers & Heard, 1997). La ecología de comunidades en combinación con las filogenias nos ofrece la posibilidad de afrontar preguntas ecológicas en un contexto evolutivo. Por último, los episodios de diversificación observados en las filogenias del *amoA* reflejan la adaptación a diferentes ambientes basados en cambios y adaptaciones a nivel molecular.

Firmas de evolución molecular en las AOA

La habilidad de las AOA a adaptarse una multitud de condiciones ambientales diferentes, convierten al *amoA* en un objetivo muy conveniente para analizar la firmas de evolución adaptativa de proteínas. Desde que la teoría neutra de la evolución molecular fue propuesta en 1960 (Freese & Yoshida, 1965; Kimura, 1968) (más tarde se formuló la teoría casi neutra (Ohta, 1992)) ha generado una gran controversia para establecer en si la selección natural modela los patrones de variación. Tras más de 40 años, todavía existe un intenso debate entre los neutralistas y los seleccionistas. Este debate va más allá del objetivo de esta tesis doctoral, pero como breve introducción para el lector, desde el 1960 coexisten dos modelos en como la evolución molecular se lleva a cabo; uno dominado por la deriva genética de la mutaciones neutras (neutralistas) y el otro enfatiza que la selección natural de mutaciones ventajosas es la fuerza más importante (seleccionistas). Después de todo parece razonable concluir que tanto la selección natural como la deriva genética determinan el destino de las mutaciones.

A lo largo de esta tesis doctoral, los diferentes tipos de selección son definidos de la misma forma utilizada en la literatura sobre evolución molecular

para evitar confusiones. Definimos selección negativa o purificadora, a cualquier tipo de selección donde las mutaciones son seleccionadas en contra; y la selección positiva es cualquier tipo de selección donde las nuevas mutaciones son ventajosas. La selección diversificadora es un tipo de selección positiva que incrementa la variabilidad y es el mecanismo clave para la divergencia de las especies; identificar las proteínas o los lugares que experimentan selección diversificadora puede de ser de ayuda para entender la función del gen.

Para entender la presión selectiva que ha modelado la variación genética, se utilizan las secuencias codificantes para proteínas y se evalúan las tasas relativas de substituciones sinónimas o no-sinónimas (Nei & Gojobori, 1986; Miyata & Yasunaga, 1980; Li et al., 1985a). Una aproximación común a este problema es estimar las tasas de substituciones no-sinónimas (dN o β) y sinónimas (dS or α).

Las mutaciones no-sinónimas pueden afectar la función de la proteína e influenciar eficacia de un organismo; mientras las mutaciones sinónimas no afectan a la secuencia de aminoácidos. Bajo el efecto de la selección negativa o purificadora, las substituciones no-sinónimas menos aptas se acumulan más despacio que la substituciones sinónimas; mientras que bajo los efectos de la selección positiva o diversificadora sucede lo contrario. Por lo tanto, la comparación de las tasas relativas de substituciones sinónimas o no-sinónimas es un concepto importante en el análisis de las secuencias codificadoras para poder proporcionar información sobre el tipo de selección que ha actuado en un conjunto de secuencias proteicas. La proporción $\omega = dN/dS$ (también conocida como β/α o K_a/K_s) se ha convertido en un estándar para medir la presión selectiva; cuando la $\omega \approx 1$ significa evolución neutra, $\omega < 1$ es selección negativa y $\omega > 1$ selección positiva.

Las estimas de dN significativamente diferentes de las de dS proporcionan evidencias de evolución no neutra. Una ventaja de esta aproximación es que se evita realizar suposiciones sobre la historia demográfica de las poblaciones, a diferencia de otros tests de neutralidad (Tajima, 1989, 1996; Fu & Li, 1993a,b; Deng & Fu, 1996; Fu, 1997; Misawa & Tajima, 1997; Fay & Wu, 2000), que comparan estimas del tamaño de la población efectiva (N_e) obtenidas de diferentes medidas de variación genética. La genética de poblaciones en bacterias (Joyce et al., 2002; Mes, 2008) todavía es una área de investigación por desarrollar. Aunque la N_e de los oxidantes del amonio puede estar en un rango similar a la de 10^9 estimada para *E. coli* (Nei & Graur, 1984) tenemos que tener cuidado de tener suficiente diversidad genética si queremos utilizar los métodos para estimar dN/dS.

Los estudios iniciales de la selección selectiva recaían en la proporción media dN/dS para la región de interés en la secuencia al nivel de DNA, ya fuese utilizando métodos basados en distancias (Li et al., 1985a; Nei &

Gojobori, 1986; Li, 1993; Pamilo & Bianchi, 1993; Comeron, 1995; Yang & Nielsen, 2000) o métodos basados en máxima verosimilitud (Goldman & Yang, 1994; Muse & Gaut, 1994). En algunos casos, estaremos más interesados en analizar las presiones selectivas restringidas a un numero pequeño de sitios (codones) de la secuencias. Para detectar las presiones selectivas en sitios específicos podemos utilizar *métodos de recuento*, *modelos de efectos de probabilidad aleatoria* (REL) o *modelos de efectos de probabilidad fija* (FEL). Una buena revisión sobre estos métodos se puede encontrar en Pond & Frost (2005)

Cuando buscamos firmas de evolución molecular en poblaciones microbianas, hay que tener en cuenta dos consideraciones importantes. Primero, la proporción dN/dS es inapropiado cuando comparamos cepas muy distantes (dS saturada con múltiples substituciones), o cepas muy próximas, en las cuales dN/dS es inflada por polimorfismos no-sinónimos segregadores (Rocha et al., 2006; Kryazhimskiy & Plotkin, 2008). Y segundo, la discordancia filogénetica causada por la recombinación afecta a los métodos de máxima verosimilitud para cuantificar la presión selectiva en los alineamientos de codones, pudiendo producir falsos positivos.

Las presiones selectivas tienen una gran influencia en la adaptación y pueden explicar parte de la diversidad genética y ecológica observada. Pero, también hay que tener en cuenta el repertorio génico del organismo para poder entender estos patrones. Aunque actualmente es fácil y barato secuenciar un genoma microbiano para realizar análisis comparativos, todavía hay el problema de obtener un cultivo puro. Por suerte, con el advenimiento de la metagenómica se ha podido solucionar en parte el problema pudiendo realizar análisis independientes a los genomas.

El uso de la metagenómica para la exploración de la oxidación del amonio en arqueas

Cada estudio metagenómico puede ser dividido en dos fases diferentes. Primero, un estudio del gen en el ambiente y la secuenciación aleatorio de todos los genes a la vez. El advenimiento de la secuenciación de nueva generación (NGS) ha permitido una revolución en la secuenciación y el análisis de metagenomas. El aumento del rendimiento y la reducción del coste de la secuenciación, unido a los avances tecnológicos han transformado el panorama de la metagenómica (Scholz et al., 2012). Y segundo, los análisis bioinformáticos para anotar los genes y determinar su relación con el ambiente. Con la anotación se puede derivar el estado metabólico de la comunidad examinando el potencial de proteínas para producir o consumir metabolitos.

Uno de los primeros estudios que han liberado las secuencias al público fue la expedición Global Ocean Sampling (GOS). Venter y colegas hicieron un muestreo desde el Atlántico noroeste a al Pacífico orienta tropical, produciendo 6.3 mil millones de pares de bases a partir de 7.7 millones de *reads* (Rusch et al., 2007). Un estudio metagenómico nos permite cubrir una extensión más amplia en las escalas de variación espacial y temporal de las comunidades microbianas marinas, por ejemplo, un muestra de GOS representa aproximadamente a un periodo de una semana en la escala temporal y en la escala geoespacial, a unos cuantos kilómetros en horizontal y a unos cuantos metros de profundidad (Fuhrman, 2009). Y lo más importante, la liberación pública de los datos ha permitido el desarrollo de múltiples aproximaciones bioinformáticas por parte de otros investigadores.

Con la cantidad de datos que se generan en los estudios metagenómicos se pueden realizar aproximaciones holísticas para el estudio de los ecosistemas marinos (Karl, 2007). Recientemente, aproximaciones basadas en la teoría de redes (Jiang et al., 2012; Faust & Raes, 2012) se han empezado a utilizar para extraer información valiosa a partir del estudio de los patrones de coocurrencia de los genes del rRNA o de los dominios proteicos involucrados en los procesos biológicos. La combinación de la metagenómica y las aproximaciones basadas en la teoría de redes ofrecen un nueva perspectiva para descubrir las piezas que faltan en los procesos biológicos como en la oxidación del amonio realizada por las AOA. En esta tesis doctoral hemos desarrollado y aplicado una aproximación basado en Modelos Gráficos Gaussianos para explorar las asociaciones funcionales de los genes y abordar uno de los principales desafíos de la metagenómica, la exploración de la fracción desconocida del universo proteico de los metagenomas.

1.2 Estructura, objetivos y resultados

El objetivo principal de esta tesis es explorar la ecología y la evolución de los microorganismos nitrificantes. La oxidación del amonio es una de las piezas clave del ciclo del nitrógeno. Tanto las bacterias como las arqueas oxidadoras del amonio se pueden encontrar coexistiendo a lo largo de diferentes ambientes. Pero cuando la primera arquea oxidadora del amonio fue aislada, se puso en relevancia la importancia de estas en comparación con las bacterias en los ciclos biogeoquímicos globales. Desde entonces hemos sido inundados por una avalancha de secuencias génicas de estas arqueas, mostrando una gran capacidad de diversificación y adaptación a ambientes diferentes.

Al no disponer de suficientes datos para realizar una aproximación holistica utilizando genómica de poblaciones y de ecología inversa para poder discernir los mecanismos ecológicos y evolutivos relacionados con la adaptación; nos hemos centrado en estudiar la secuencia del *amoA*. La amonio monooxigenasa es la enzima responsable de la oxidación del amonio, para su estudio hemos aplicado una combinación de técnicas de ecología de comunidades y de evolución molecular con el objetivo de entender los mecanismos de los patrones de diversificación observados.

Por otra banda, otro de los misterios asociados a la oxidación del amonio por parte de las arqueas, es su inusual bioquímica para realizar la oxidación del amonio. En arqueas faltan todos los elementos necesarios para llevar a cabo la oxidación del amonio a excepción del AMO. Para poder aportar algo de luz a este misterio hemos desarrollado un potente método basado en modelos gráficos para capturar todas las asociaciones funcionales presentes en los metagenomas basado en sus co-ocurrencias ecológicas.

Los objetivos detallados y la estructura de esta tesis están detallados a continuación:

Capítulo 4: Filogenias de comunidad de las oxidadoras microbianas del amonio

Los microorganismos que median la oxidación del amonio desempeñan un papel fundamental en la conexión entre la fijación biológica de nitrógeno y las pérdidas anaeróbicas de nitrógeno. Las bacterias y arqueas oxidantes del amonio (AOB y AOA, respectivamente) han colonizado ambientes similares alrededor del mundo. La oxidación del amonio es la etapa limitante en la nitrificación, y la amonio monooxigenasa (AMO) es la enzima clave para este proceso. La ecología molecular de la oxidación del amonio ha sido ampliamente explorada mediante estudios de la subunidad A del gen AMO (*amoA*). En este estudio, hemos explorado la ecología de comunidades de las AOB y las AOA, analizando 5776 secuencias génicas del *amoA* aisladas de más de 300

lugares diferentes, y clasificando los hábitats utilizando ontologías ambientales. En resumen, la riqueza filogenética es mayor en AOA que en AOB, y los sedimentos contienen la mayor riqueza filogenética mientras que el plancton marino la más baja. También se ha observado que los oxidantes del amonio de agua dulce son filogenéticamente más ricos que sus homólogos marinos. Las comunidades de AOA son más diferentes entre sí que las de AOB y se observan linajes monofiléticos para los sedimentos, suelos y plancton marinos para AOA pero no para AOB. Los patrones de diversificación muestran una cladogénesis más constante a través del tiempo para las AOB mientras que en AOA se suceden dos eventos de diversificación rápidos separados por un largo episodio de no diversificación. Los índices de diversificación (γ estadístico) para la mayoría de los hábitats indican un $\gamma_{AOA} > \gamma_{AOB}$. Los suelos y sedimentos experimentan estallidos de diversificación tempranos mientras que los hábitats generalmente eutróficos y ricos en amonio, como las aguas residuales y lodos, muestran una aceleración en las tasas de diversificación en el presente. En general, este trabajo muestra por primera vez una visión global de la estructura de la filogenia de comunidades de los dos grupos de oxidadoras del amonio (AOA y AOB), siguiendo los más estrictos estándares para su análisis. y proporciona un punto de vista ecológico en los caminos diferenciales evolutivos experimentados por la generalización de oxidantes de amoníaco microorganismos. La imagen que se obtiene de la distribución AOB y AOA en los diferentes hábitats proporciona un nuevo punto de vista para comprender la ecofisiología de los oxidantes del amonio en la Tierra.

Capítulo 5: Patrones evolutivos de las arqueas oxidadoras de amonio

La *amoA* es uno de los genes clave implicados en la oxidación del amonio y presenta una distribución muy diversa y abundante en los diferentes ambientes del planeta. En el siguiente estudio, analizamos los procesos evolutivos moleculares implicados en la alta capacidad diversificadora de este gen; aunque el gen de la *amoA* se encuentra bajo los efectos de la selección purificadora, hemos encontrado evidencias de selección episódica diversificadora en codones individuales así como en linajes. Hemos observado eventos de selección postiva diversificadora seguido de periodos de conservación (selección homogeneizante) como un mecanismo para la generación y mantenimiento de un *seed bank* evolutivo del gen del *amoA* como un mecanismo para la radiación adaptativa en los diferentes hábitats.

Capítulo 6: Análisis de la distribución de AOA por marcaje en ambientes marinos

Los Polimorfismos de longitud de fragmentos de restricción (T-RFLP) es una técnica utilizada para analizar comunidades microbianas complejas. Permite la cuantificación de los filotipos más abundantes y se ha utilizado principalmente para comparar diferentes comunidades. T-RFPred ha sido desarrollado para identificar y asignar información taxonómica a los picos de los cromatogramos obtenidos en los T-RFLP, para poder realizar una descripción más intensiva de las comunidad microbianas. El programa estima el tamaño esperado de los 16S rRNA representativos para un determinado cebador y enzima de restricción y proporciona una asignación taxonómica.

Capítulo 7: Una visión ecológica y evolutiva a la distribución espacial de los *Thaumarchaeota* marinos

Los *Thaumarchaeota* marinos son unos contribuyentes importantes en las primeras etapas de la nitrificación en los oceanos y además tienen un importante papel en les ciclos biogequímicos marinos. Estos *Thaumarchaeota* marinos presentan dos grupos diferentes correspondientes a un ecotipo de superficie (*shallow*) y otro de aguas profundas (*deep*). Esta partición observada se cree que es originada por adapataciones relacionadas con procesos de fotoinhibición-resistencia y adaptación a la presión hidrostática. En el siguiente trabajo analizamos los diferentes ecotipos para determinar las presiones selectivas que han actuado sobre el gen del *amoA*. Hemos encontrado al codon {89} como un potencial componente clave en las adaptaciones a la luz. Por otro lado, hemos encontrado evidencias que el gen del *amoA* se encuentra bajo intensa selección a nivel molecular para adaptarse a las condiciones de las profundidades marinas. Por último, hemos encontrado que el linaje donde donde diversifican los dos ecotipos están sujeto a selección episódica diversificadora; posiblemente como resultado de la rápida diversificación y radiación adaptativa que este grupo ha experimentado en la columna de agua marina.

Capítulo 8: Análisis de redes para la exploración de la oxidación del amonio en metagenomas marinos

El análisis del genoma de Candidatus *Nitrosopumilus maritimus* SCM1, un *Thaumarchaeota* marino, ha revelado la existencia de un sistema diferente para la oxidación del amonio del descrito para bacterias. Actualmente se barajan dos hipótesis para explicar la falta del homólogo para la hydroxilamina oxidasa encontrado en bacterias. Unos sugieren que el nytroxyl puede ser utilizado como intermediario en lugar de la hidroxilamina; y otros que el proceso

puede ser mediado por oxidasas presentes en el espacio periplásmico. En el siguiente trabajo hemos aplicado Modelos Gráficos Gaussianos para analizar la oxidación del amonio en arqueas. Hemos combinado el conocimiento de las familias de proteínas conocidas con la fracción proteica desconocida para encontrar asociaciones en términos de función y estructura. Hemos sido capaces de determinar estas asociaciones a los genes de *Nitrosopumilus* y generar una serie de candidatos que podrían estar implicados en la oxidación del amonio.

1.3 Conclusiones

Las conclusiones generales de esta tesis son las siguientes:

- AOA y AOB presentan estructuras de comunidad muy diferentes en términos de riqueza filogenética y β-diversidad.
- Los patrones de diversificación muestran una cladogénesis más constante en el tiempo para las AOB mientras que las AOA han sufrido dos eventos de diversificación separados por un largo estado de estasis. La tasa de diversificación (estadístico γ) para la mayoría de hábitats indican que $\gamma_{AOA} > \gamma_{AOB}$
- La combinación de métodos evolutivos basados en codones y los de filogenias de comunidades resultan una herramienta valiosa para entender los procesos de diversificación de los genes marcadores ambientales
- El gen de la *amoA* muestra evidencias de selección purificadora debido a sus restricciones funcionales. Aunque, hay evidencias de selección episódica diversificadora tanto para codones individuales como para linajes.
- Los patrones de diversificación de la *amoA* muestran evidencias de la generación y mantenimiento de un *seed bank* evolutivo que dirige la radiación adaptativa en los diferentes grupos filogenéticos.
- El gen de la *amoA* en los océanos muestra patrones de selección episódica diversificadora para las diferentes condiciones ambientales de los ecotipos *shallow* y *deep*. La *amoA* de la OMZ se encuentra sujeto a fuertes constricciones funcionales.
- Un enfoque pionero combinando familias de *desconocidos* con Modelos Gráficos Gaussianos para analizar las asociaciones funcionales en los metagenomas es un herramienta valiosa para realizar predicciones funcionales.

Informe del director

El **Dr. Emilio Ortega Casamayor**, Investigador Científico del Centro de Estudios Avanzados de Blanes (CEAB) perteneciente al Consejo Superior de Investigaciones Científicas (CSIC), en calidad de Director de la tesis doctoral *Aproximación Genética a la Ecología y Evolución de Microorganismos Nitrificantes* presentada por Antoni Fernàndez Guerra para optar al título de Doctor dentro del programa de doctorado de Genética de la Universidad de Barcelona, hace constar que la participación del aspirante a doctor en cada uno de los artículos presentados en esta memoria es la que se detalla en los párrafos siguientes. Así mismo constata que esta tesis está formada por 2 artículos científicos del SCI ya publicados y 3 que se encuentran en fase de publicación. En todos ellos, el candidato a doctor es primer firmante de los trabajos y en ningún caso el resto de coautores ha utilizado, implícitamente o explícitamente, datos o resultados de estos trabajos para la elaboración de ninguna otra tesis doctoral.

- **Artículo I**
  Fernàndez-Guerra A, EO Casamayor (2012) Habitat-Associated Phylogenetic Community Patterns of Microbial Ammonia Oxidizers. *Plos One* 7 (10): e47330.

  Indicadores bibliométricos. IF (2011): 4.092; Cuartil: Q1; Categoria: BIOLOGY (posición 12 de 85). El candidato a doctor lideró desde el primer momento el diseño experimental, la aplicación de metodologías y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo. En sus dos primeros meses, el trabajo ha sobrepasado las 600 visitas y las 150 descargas del pdf.

- **Artículo II**
  Fernàndez-Guerra A, EO Casamayor (manuscrito en preparación) Evolutionary patterns in archaeal ammonia oxidizers.

  El candidato a doctor lideró desde el primer momento el diseño experimental, la aplicación de metodologías y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo.

- **Artículo III**
  Fernàndez-Guerra A, Buchan A, Mou X, EO Casamayor, González, JM

(2010) T-RFPred: a nucleotide sequence size prediction tool for microbial community description based on terminal-restriction fragment length polymorphism chromatograms *BMC Microbiology* 10:262-269.

Indicadores bibliométricos. IF (2011): 3.044; Cuartil: Q2; Categoria: MICROBIOLOGY (posición 43 de 114). El candidato a doctor participó en el diseño experimental, y llevó a cabo la obtención de los datos, la aplicación de metodologías de análisis filogenético y la redacción del primer borrador del manuscrito. El método ha sido aplicado satisfactoriamente para el estudio de la distribución de archaea oxidadoras de amonio en ambientes marinos.

- **Artículo IV**
  Fernàndez-Guerra A, EO Casamayor (manuscrito en preparación) An evolutionary perspective on the phylogenetic partitioning in marine ammonia-oxidizing *Thaumarchaeota*.

  El candidato a doctor lideró desde el primer momento el diseño experimental, la aplicación de metodologías y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo.

- **Artículo V**
  Fernàndez-Guerra, A, A Barberán, R Kottmann, FO Glöckner, EO Casamayor (manuscrito en preparación) A network approach to explore the archaeal ammonia oxidation in oceans trough metagenomics.

  El candidato a doctor lideró desde el primer momento el diseño experimental, la aplicación de metodologías y la redacción del primer borrador del manuscrito, siendo primer autor del trabajo.

El director de la tesis
Dr. Emilio Ortega Casamayor
Inv. CEAB-CSIC

Ecology and Evolution of Microbial Nitrifiers

2
General Introduction

Nitrogen (N) is one of the most important elements for life in Earth, many of the key compounds in biochemical processes required for live include N as a critical component, nucleotides or aminoacids are good examples. Nitrogen gas (N_2) comprises around the 80% atmosphere, but due the triple bond of this gas such N reservoir is not biologically available for most organisms on Earth. The triple bond has to be broken, but this is a very energy-demand reaction. Nitrogen fixation is the process of breaking the triple bond to obtain reactive nitrogen (Nr) . Although N fixation can be carried out by natural processes like lightning or in high-temperature circulation systems (Canfield et al., 2006), the amount produced is not enough to supply the necessary Nr for life developed along the Earth's history. Most of the Nr proceed from the fixation mediated by certain microorganisms that have developed an special metabolic machinery to break the triple-bonded N_2 and generate biologically active reduced forms of nitrogen that later could be assimilated by other organisms.

For millions of years Nr was originated by microbial nitrogen fixation processes being the availability of Nr in the biosphere a key growth limiting factor. Such limitation derived in a competition between the different life forms shaping the actual biodiversity and the internal relationships among its members. This scenario totally changed after chemists discovered the essential role of nitrogen in life biochemistry, and other scientists identified it as an essential nutrient for plants and animals (Smil, 2004). In parallel, there were an increase concern about the growing food demand by the human population, being the estimated growth rate exceeding the known sources by far. This limitation scenario changed after the biological fixation of N was unveiled. Once the basis for the natural process of Nr were settled, the invention of the chemical pro-

cess to convert atmospheric N_2 to NH_3 was the next step. In 1913 the Haber-Bosch process was invented, and for the first time in the history of Earth, there was the possibility to have unlimited supplies of Nr; was the beginning of artificial fertilizers. There is a direct correlation with the population growth and the discovery of the Haber-Bosch process as seen in Figure 2.1. But the Haber-Bosch process wasn't the only anthropogenic source of Nr. Since the industrial revolution, energy production from fossil fuel combustion directly injects Nr into the atmosphere.

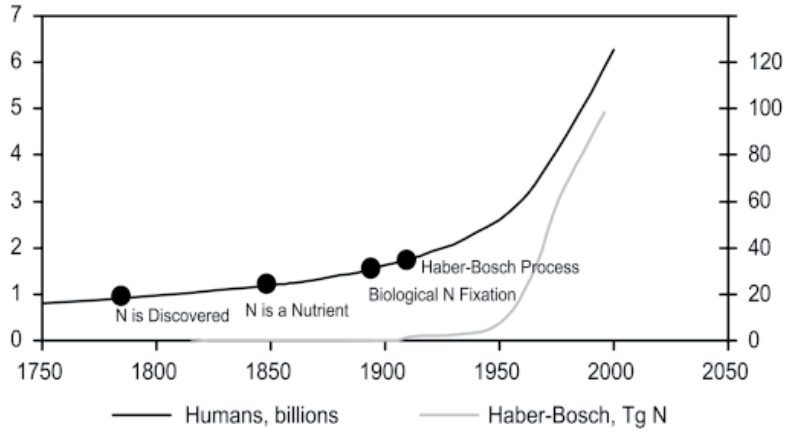

Figure 2.1: Global population trends (Y scale) with key dates for the discovery of N as an element in the periodic table and its role in various biogeochemical processes. The estimated annual production of Nr by the Haber-Bosch process is also shown (Y2 scale). Figure from Galloway & Cowling (2002)

Human activity has totally unbalanced the N cycle and for the first time in Earth's history Nr is in excess. The rate of Nr generation excedes the conversion back to N_2 by denitrification processes and Nr accumulates in the environment. The accumulation of Nr in nature is a serious issue with negative consequences to humans and ecosystems, ranging from to eutrophication and acidification of terrestrial and aquatic systems to the loss of stratospheric ozone, among others (Galloway & Cowling, 2002). However, the main concern on the excess of Nr is what is known by *nitrogen cascade*.From Gruber & Galloway (2008):

> "For example, an emitted molecule of nitrogen oxide can first cause photochemical smog and then, after it has been oxidized in the atmosphere to nitric acid and deposited on the ground, can lead to ecosystem acidification and eutrophication."

But the complexity of the problem increases when we try to understand

how this change in the nitrogen cycle affects to the other biogeochemical cycles (Figure 2.2), and particular with the carbon cycle, as they are directly linked. On the one hand, anthropogenic factors are related with the constant increase of the availability of Nr and the CO_2 atmospheric concentration, one of the key players of global warming. There is a direct relationship between artificial fertilization of soils and CO_2 uptake from intensive agriculture crops, lowering the levels of atmospheric CO_2 and increasing the vegetal biomass (Schimel et al., 2001). And on the other hand, the nitrogen and carbon cycle are tightly related as a result of the life processes. Nitrogen, carbon, phosphorous and other elements are used for building the basic blocks of life that will result on the myriad of life forms that inhabits the Earth planet. The coupling between the different elements occurs at specific stoichiometries that determines the linking of the different biogeochemical cycles (Sterner & Elser, 2002). For instance, the C/N ratio is different in marine or terrestrial photosynthetic organisms, while in the marines there is a small fluctuation in the ratio, in the terrestrials, the variation is higher.

The amount of Nr on Earth is controlled by the biological fixation and denitrification processes, but as shown in Figure 2.2, alterations to those processes can affect to the global carbon cycle and climate, meanwhile the C/N ratio of autotrophs remains immutable. Figure 2.2 shows the differences in the N cycle and the interaction with the other biogeochemichal cycle between terrestrial and marine habitats. Terrestrial ecosystems are more affected by anthropogenic perturbations while in marine ecosystems this effect is negligible.

There is not a full agreement about how biological fixation and denitrification are balanced in oceans (Gruber & Galloway, 2008; Codispoti, 2007) but it is well established both that the marine nitrogen cycle is very dynamic, the Nr turn over is less than 3,000 years (Gruber & Galloway, 2008); and the marine phosphorous cycle is the keystone for stabilizing the marine nitrogen cycle (Deutsch et al., 2007). Terrestrial systems are not so well studied, but there is a strong lateral transport from land, where the sources of Nr exceeds the denitrification, into freshwater systems, where denitrification prevails. In fact, in terrestrial systems, nearly half of the denitrification occurs in freshwater systems (Seitzinger et al., 2006).

The processes of fixation and desnitrification are mainly driven by specific groups of microorganisms. With all the perturbations introduced in the global N cycle by recent human activities and the changes induced at the global scale, the study and understanding of microbial communities directly linked to the N cycling has became a hot topic in microbial ecology. Nowadays understanding and expanding the knowledge on the microbial processes implicated in nitrification-denitrification are crucial in terms of maintaining the balance between biogeochemical cycles. One of the limiting steps in natural denitri-

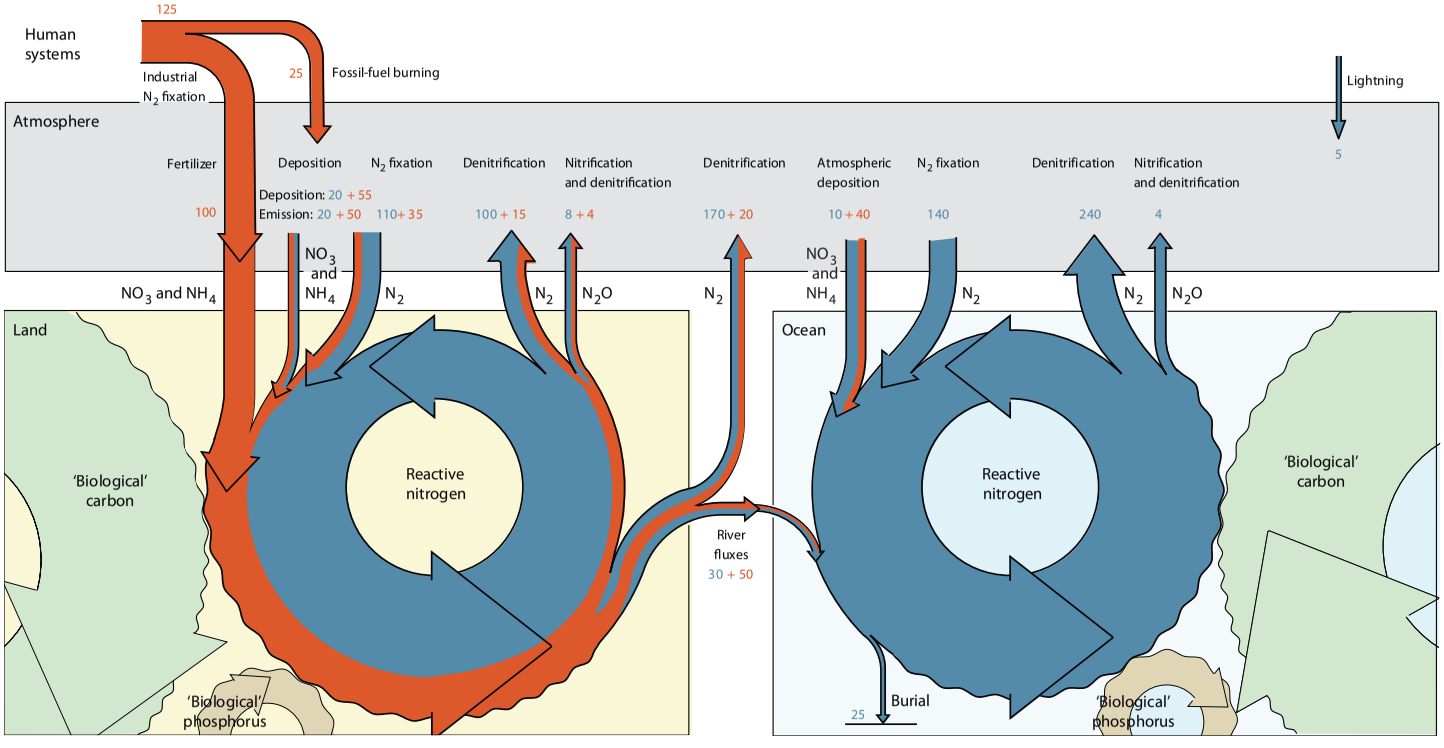

Figure 2.2: Summary of the processes that transform molecular nitrogen into reactive nitrogen, and back. There is also shown the other biogeochemical cycles (carbon and phosphorus coupled to the N cycle on land and oceans. Blue fluxes denote 'natural' (unperturbed) fluxes; orange fluxes denote anthropogenic perturbation.The numbers(in Tg N per year) are values for the 1990s. Figure from Gruber & Galloway (2008)

fication is the continuous feeding with nitrate through nitrification. In the nitrification reaction, the NH_3 provided from artificial or natural nitrogen fixation and organic or atmospheric sources, is converted to nitrate via nitrite, that could be further used by denitrifiers or assimilated by the different organisms. Nitrification is carried out in two steps by two physiologically distinct groups of autotrophic organisms: ammonia oxidizers for NH_3 to NO_2^-, and nitrite-oxidizers, from NO_2^- to NO_3^-.

Ammonia oxidation (Figure 2.3) is the rate-limiting step of nitrification and is a biogeochemical process of global importance in natural and artificial ecosystems worldwide having a substantial environmental impact on greenhouse gas emissions (mainly nitrous oxide N_2O, and nitrogen oxides NOx) through nitrification–denitrification processes.

Ammonium concentrations in many soils have increased in recent years as a result of land-use changes and increases in atmospheric ammonium concentrations (Rockström et al., 2009). Those global changes related to the nitrogen cycle may have influenced the microbial ecology of the nitrification process(Verhamme et al., 2011) in soils, where ammonia oxidation can lead to significant net nitrogen loss through subsequent denitrification or leaching of nitrate. In marine environment, nitrification accounts for about half of the nitrate consumed by growing phytoplankton at the global scale (Yool et al., 2007) and is responsible for the deep ocean nitrate reservoir (Karl, 2007), the largest pool of reactive nitrogen in the biosphere (Beman et al., 2010; Leininger et al., 2006). Also in aquatic environments, ammonia oxidation is an important component of nitrogen mineralization from organic sources and for the removal of anthropogenic nitrogen inputs in coastal waters. In artificial ecosystems such as sludge, bioreactors or wastewater treatment plants, ammonia oxidation is also a key step in the removal of nitrogen (Van Loosdrecht & Jetten, 1998; Mussmann et al., 2011). In freshwater ecosystems, nitrification could remove excessive ammonium nitrogen and prevent lakes from eutrophication (Hagopian & Riley, 1998).

Therefore, microbial ammonia oxidizers play a fundamental role in the connection between biological N fixation and anaerobic N losses, and are widely detected in a large variety of aquatic and terrestrial environments (Nicol & Schleper, 2006; Auguet et al., 2010). Microbial ammonia oxidizers are found in Bacteria (AOB) and Archaea (AOA) domains. AOB and AOA have colonized similar worldwide distributed environments but with different degrees of success in abundance, activity, and distribution(Prosser & Nicol, 2008; Martens-Habbena et al., 2009; Auguet et al., 2011; Sauder et al., 2011; Verhamme et al., 2011).

The molecular ecology of the ammonia oxidation process has been extensively explored surveying the subunit A of the ammonia monooxygenase

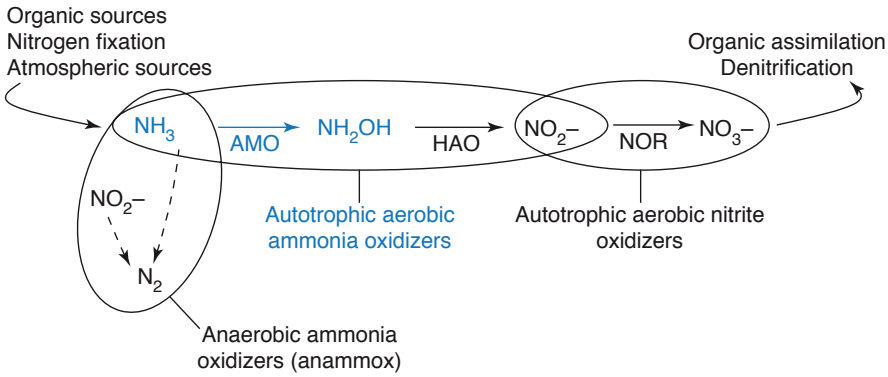

Figure 2.3: Autotrophic ammonia oxidation during nitrification. Ammonia-oxidising organisms convert ammonia to nitrite through hydroxylamine using ammonia monooxygenase (AMO) and hydroxylamine oxidoreductase (HAO). Autotrophic nitrite oxidizers subsequently use the enzyme nitrite oxidoreductase (NOR) to convert nitrite to nitrate, which can be assimilated or subjected to denitrification processes. In anaerobic environments, ammonia can be converted to molecular nitrogen by the 'anammox' process by several enzymatic steps (represented by dashed arrows).Figure modified from Nicol & Schleper (2006)

(AMO) (Auguet et al., 2011; Rotthauwe et al., 1997; Agogué et al., 2008; Francis et al., 2005, 2007). Both AOA and AOB contain the *amoA* gene encoding the alpha subunit of the AMO; however, the gene sequence has evolved separately in each of the phylogenetically distinct but physiologically related ammonia oxidizers microorganisms (Treusch et al., 2005). The AMO enzyme is composed of three subunits (gene products of *amoA*, *amoB* and *amoC*) and is evolutionary and functionally related to particulate methane monooxygenase (pMMO) enzymes of methane-oxidising bacteria (Holmes et al., 1995).

2.1 Bacterial ammonia oxidizers

Microbial ammonia oxidation was initially considered to be restricted to a few bacteria, specifically within the phylum Proteobacteria (γ and β), which under laboratory conditions mostly show an affinity threshold for ammonium higher than the concentrations usually found in situ (Bollmann et al., 2002).

For many years environmental molecular studies of AOB have targeted 16S rDNA or rRNA of β-subclass assuming that members of the γ-subclass were absent or insignificant as γ-subclass ammonia oxidizers had been only detected as isolates only from marine environments. That approach was totally biased as they relied on culture-based methods, but the reality was that the majority of environmental microorganisms were not growing into pure

cultures on standard laboratory media. Then in 1993 McTavish et al. (1993) published the first *amoA* gene sequence from the cultured AOB *Nitrosomonas europaea*. Since then, the α-subunit of *amoA*, was used a marker and the number of sequences in the databases associated to the *amoA* gene increased rapidly.

The use of the *amoA* gene as a marker instead of the rRNA, provided some advantages. First, *amoA* genes from β- and γ-AOBs can be amplified at the same time (Holmes et al., 1995; Mendum et al., 1999; Nold et al., 2000; Sinigalliano et al., 1995). Second, the *amoA* gene shares and an evolutionary ancestor with the gene encoding the α-subunit of particulate methane monooxygenase (*pmoA*) (Holmes et al., 1995; Nold et al., 2000). Sufficient sequence homology remains to allow both *amoA* and *pmoA* genes to be targeted simultaneously, allowing an assessment of the relative abundances of autotrophic ammonia and methane oxidizing bacteria in situ.

The *amoA* genes of β-subgroup ammonia oxidizers can also be targeted specifically using primers that do not detect γ-proteobacterial *amoA* or *pmoA* when only *beta*-subclass ammonia oxidizers are under study (Rotthauwe et al., 1997; Stephen et al., 1999). Furthermore, *amoA* contains a greater level of sequence variation than 16S rDNA, providing better resolution of closely related strains.

Bacterial ammonia oxidizers are usually found in nitrogen rich environments such as sewage, bioreactors, biofilms and biofilters. AOB are found in terrestrial and aquatic environments (freshwater and marine). But, for many years, microbial ecologists remained perplexed with the nitrifying capacity of many ecosystems where apparently bacterial ammonia oxidizers were far below detection limits, particularly under the most oligotrophic conditions (Olson, 1981).

2.2 Archaeal ammonia oxidizers

Early in the 1990s, as part of a census of bacterial diversity in the nitrifying reactor systems used to treat water in a saltwater aquaria at the Shedd Aquarium (Chicago, Illinois), it was hypothesized that a novel group of ammonia oxidizers may exist (Fuhrman et al., 1992) because PCR-amplifications targeting known bacterial ammonia oxidizers failed to obtain the targeted organisms. A few years later, equivalent observations were made in a study of the microbiology of nitrogen processing in Plum Island Sound (Massachusetts) estuary sediments (Könneke et al., 2005) and again specific primers for bacterial ammonia oxidizers failed to amplify the expected sequences. However, this study was expanded to archaeal 16S rDNA gene sequences from a marine clade known as marine Group I *Crenarchaeota*. This clade was found to be

abundant in estuary sediment enrichments and in nitrifying filtration systems at the Shedd Aquarium (Könneke et al., 2005). Those results suggested a close relationship between the abundant Group I *Crenarchaeota* and ammonia oxidation. The next step was to obtain enriched cultures using saltwater aquaria material as inoculant, and a growth medium containing a much lower concentration of ammonia than typically used to enrich bacterial ammonia oxidizers. Finally an active ammonia-oxidizing culture highly enriched in an archaeal population affiliated with the Group I *Crenarchaeota* was obtained, and the first pure culture of an AOA, *Nitrosopumilus maritimus* strain SCM1 was described [nitrosus (Latin): nitrous; pumilus (Latin): dwarf; maritimus (Latin): of the sea].

The combination of culture-independent methods to associate the ammonia oxidation to the marine Group I *Crenarcheota* and the further isolation in culture, fuelled the culture of other AOA from different environments, such as, *Nitrososphaera viennensis* [nitrosus (Latin): nitrous (nitrite producer); sphaera (Latin): spherically shaped; viennensis (Latin): from Vienna] was isolated from soil (Tourna et al., 2011) and *Nitrosotalea devanaterra* [nitrosus (Latin): nitrous (nitrite producer); talea (Latin): slender rod; devana (Latin): Aberdeen; terra (Latin): soil], from a low pH soil (Lehtovirta-Morley et al., 2011). Since the isolation of *Nitrosopumilus maritimus* and the other AOA, the phylogenetic diversity of the ammonia oxidizing archaea has been revisited several times while new organisms are being described (Pester et al., 2012). Nowadays, AOA phylogenies are built using large datasets of 16S rDNA genes and *amoA* α-subunit, and have revealed new phylogenetic relationships like the *Thaumarchaeota*, a new division within Archaea where all characterized AOA (Brochier-Armanet et al., 2008; Pester et al., 2011; Spang et al., 2010) are found. A common characteristic for the cultured *Thaumarchaeota* is the ability to grow via ammonia oxidation, as *Thaumarchaeota* are a deeply diverging group within the Archaea, new questions arises for the origins of archaea and the nitrogen cycle because the anaerobic physiology dominates the deep clades in Archaea.

The environmental distribution of AOA is extremely diverse exceeding their bacterial counterparts. Abundance and diversity patterns have been inferred primarily from sequencing and quantification of the gene coding for the putative archaeal amoA (Francis et al., 2005; Beman et al., 2008). Ammonia moonoxygenase subunit A is usually found well correlated with the abundance of crenarchaeol (Damsté et al., 2002; Könneke et al., 2005; Leininger et al., 2006). Results from Lehtovirta et al. (2009); Ochsenreiter et al. (2003) and Leininger et al. (2006) suggest that AOA appear to be the dominant archaeal clade in soils, comprising 1-5% of all prokaryotes and are a dominant marine group, comprising 20-40% of all marine bacterioplankton (Karner

et al., 2001; Church et al., 2003). AOA also appear to be a major ammonia-oxidizing population in geothermal habitats (de la Torre et al., 2008; Zhang et al., 2008; Reigstad et al., 2008). Cultured AOA can grow at temperatures as high as 74ºC (*Nitrosocaldus yellowstonii*). This broad habitat range in the AOA is also reflected by the tremendous phylogenetic diversity defined by 16S rRNA sequence types recognized to have ammonia-oxidizing affiliates and the corresponding diversity of putative archaeal *amoA* sequences.

2.3 Differences between AOB and AOA: two different ammonia oxidizing strategies

Martens-Habbena & Stahl (2011) compared the stoichiometry of ammonia oxidation between *N. maritimus* and AOB using microrespirometry to measure oxygen and ammonia consumption relative to nitrite and they concluded that the overall stoichiometry in AOA is indistinguishable from that of AOB:

$$1NH_3 + 1.5O_2 \Rightarrow 1NO_2^- + H_2O + H^+$$

In Bacteria, the ammonia monooxygenase is the key functional enzyme responsible for the conversion of ammonia to hydroxylamine, which is subsequently converted to nitrite by hydroxylamine oxidoreductase; but, the genome analysis of the marine *Thaumarchaeota* Candidatus *Nitrosopumilus maritimus* SCM1, revealed the existence of an ammonia oxidation system different from the one found in AOB. Thus, two hypotheses were proposed to explain the lack of the homolog for the hydroxylamine oxidase complex and the capacity for synthesis of c-type cytochromes(Figure 2.4). One hypothesis assumes that it could may driven by one of the periplasmic multicopper oxidases; and the other suggests the use of nytroxyl as intermediate instead of hydroxylamine (Walker et al., 2010).

In addition, only two small plastocyanin-like proteins are shared by all AOA revealed after comparative genomic analyses (Stahl & de la Torre, 2012). In the nytroxyl hypothesis, these copper containing proteins may participate in electron transfer from nitroxyl to a membrane-bound electron transfer chain (Figure 2.4). Although the archaeal pathway for ammonia oxidation has not been resolved by comparative genomics, recent studies using nitric oxide (NO) sensitive microelectrodes indicate that NO may function in the biochemistry (Stahl & de la Torre, 2012). Measurable amounts of NO are produced during ammonia oxidation. This has led to hypothesize that NO may be an intermediate or function as a redox shuttle, for instance, delivering electrons to the AMO (Figure 2.4).

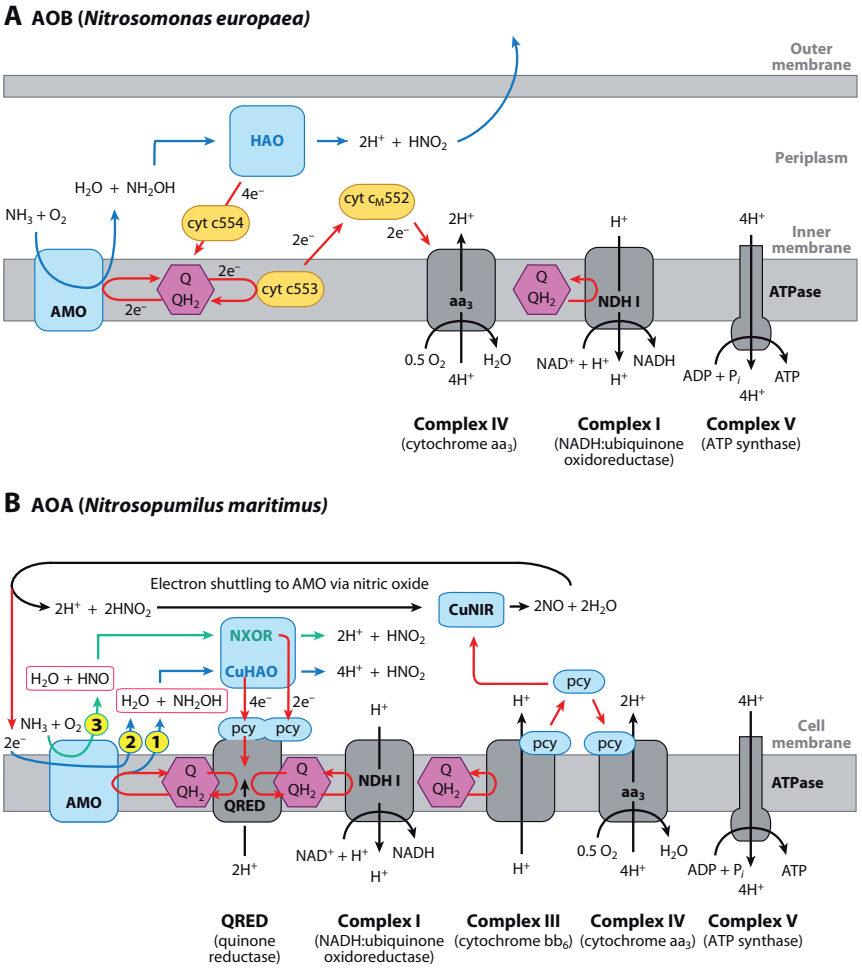

Figure 2.4: (A) Proposed pathway for ammonia oxidation in the AOB *Nitrosomonas europaea*. Ammonia is oxidized to NH_2OH by the membrane enzyme complex AMO. Subsequently, hydroxylamine is oxidized to nitrite in the periplasm by HAO. (B) Proposed pathway for ammonia oxidation in the AOA *Nitrosopumilus maritimus*. There are three alternative pathways for the ammonia oxidation. Pathway 1 is of the bacterial type, in which electrons produced by the oxidation of hydroxylamine to nitrite by a presumed CuHAO are transferred to pcy electron carriers and then to the quinone pool by a membrane-associated QRED. Pathway 2 speculates that NO, produced by the reduction of nitrite by a proposed CuNIR, is the source of electrons for AMO. The possibility that HNO is the product of the archaeal AMO is shown by pathway 3. This pathway would eliminate the requirement for electron recycling during the initial oxidation of ammonia. Subsequently, HNO would be oxidized to nitrite by a presumed NXOR.
Red arrows indicate electron flow. Blue shading denotes copper-containing proteins. Hexagons containing Q and QH_2 represent the oxidized and reduced quinone pool, respectively. Abbreviations: AOA, ammonia-oxidizing archaea; AOB, ammonia-oxidizing bacteria; AMO, ammonia monooxygenase; HAO, hydroxylamine oxidoreductase; NO, nitric oxide; HNO, nitroxyl; CuHAO, copper hydroxylamine oxidoreductase; CuNIR, copper-dependent nitrite reductase; NXOR, putative nitroxyl oxidoreductase; pcy, plastocyanins; NDH, NAD(P)H:quinone oxidoreductase; NH_2OH, hydroxylamine; QRED, quinone reductase. Figure and legend from Stahl & de la Torre (2012).

In contrast, the AOB take the electrons required by the monooxygenase from the membrane-associated quinone pool. If in AOA the formation of nitroxyl as the first product of ammonia oxidation or the use of NO as an electron redox shuttle for hydroxylamine generation would eliminate the drawing of electrons directly from the quinone pool, either by obviating a requirement for reductant through formation of nitroxyl or by drawing electrons from a lower potential donor in the reduction of nitrite to NO (Figure 2.4B). Pathways 1 and 2 in Figure 2.4B show possible recycling of an NO redox shuttle. The associated thermodynamic calculations assume that electrons for nitrite reduction originate from a donor species with an electrical potential (230 mV) approximately that of a c1-type cytochrome and in the known range of plastocyanins.

Because the formation of hydroxylamine as the immediate product of the presumptive AMO has not been demonstrated yet, the archaeal pathway for ammonia oxidation must be considered unresolved at this time.

2.4 Phylogenetic ecology of microbial nitrifiers

The ecological and physiological differences and similarities between AOA and AOB make the perfect scenario to apply phylogenetic comparative methods to analyse community structure and diversification patterns using the *amoA* gene sequence. Since the development of the phylogenetic independent contrast (PIC) (Felsenstein, 1985; Ackerly, 2009) it has been a revolution in the community ecology field, with succesful but also controversial applications of phylogenies in ecological analyses (Webb et al., 2002). Ecology and evolution are intimately associated and several methodologies have been developed to analyse the evolutionary constrains observed in nature. At present, community ecologists are able to evaluate where most of the biological diversity accumulates (Faith, 1992) and how this diversity is structured (Webb, 2000; Helmus et al., 2007); or how phylogenetic β-diversity (i.e. similarity among communities based on evolutionary history) is distributed along environmental gradients (Lozupone & Knight, 2005; Bryant et al., 2008; Ives & Helmus, 2010).

The combination of phylogenetic information with traditional ecological metrics such as species richness, derived in a more complete measure, the Phylogenetic diversity (PD) (Faith, 1992). PD measures the diversity of lineages as the sum of the branch lengths from the members within a community, adding to the count of species richness the evolutionary history. The higher the PD value is the more distantly related species are in a community. Metrics like the mean pairwise distance (MPD), that measures the mean phylogenetic distance between all members from each community, or the mean nearest taxon

distance (MNTD) that calculates the mean distance separating each member from its closest relative (Webb, 2000; Webb et al., 2002) are useful to examine the ecological and evolutionary drivers of community assembly. On the other hand, the phylogenetic species variability index (PSV) quantifies how phylogenetic relatedness decreases the variance of a hypothetical neutral trait shared by all members of a community. The value is 1 when all species are unrelated (i.e. a star phylogeny) and approaches 0 as species become more related (Helmus et al., 2007). Measures of β-diversity or community similarity can be defined as the fraction of branch length shared between two communities (Lozupone & Knight, 2005; Bryant et al., 2008). Taxonomic or compositional β-diversity and phylogenetic β-diversity would be exactly the same between two communities if every species were equally related to every other one, that is, a star phylogeny.

Phylogenies derived from molecular data provide an indirect record of the speciation events that have led to extant species, reflecting the tempo and mode of microevolutionary and macroevolutionary processes related to diversification (Mooers & Heard, 1997). Community ecology in combination with phylogenies, opens the possibility to address ecological questions in an evolutionary context. Ultimately, the diversification events observed in *amoA* phylogenies reflects the adaptation to different evolutionary environments relying on changes and adaptations at the molecular level.

2.5 Signatures of molecular evolution in AOA

The ability of AOA to undergo adaptations to a large set of environmental conditions converts the *amoA* in a very convenient target to analyze the signatures of adaptive protein evolution. Since the neutral theory of molecular evolution was proposed in the 1960s (Freese & Yoshida, 1965; Kimura, 1968) (and later the nearly neutral theory (Ohta, 1992)) it has been a high controversy to establish whether or not natural selection shapes the patterns of variation. This controversy is partially caused by Kimura's definition of neutrality, which was too strict. After more than 40 years, there is still an intense debate between neutralists and selectionists. This debate goes beyond the aim of this PhD thesis, but to introduce the reader a bit into the subject, since the 1960s two conflicting models of how molecular evolution takes place coexist: one dominated by the genetic drift of neutral mutations (neutralists) and the other emphasizing that natural selection of advantageous mutations is the more important force (selectionists). Overall, it seems reasonable to conclude that both natural selection and genetic drift determine the fate of mutations. And currently the debate is more focused on finding general laws or patterns of molecular evolution for the description of particular examples where natural selection has

shaped the observed variation (Wagner, 2008; Nei, 2005).

Along this PhD thesis the different types of selection will be defined in the same way as it is used in the molecular evolution literature to avoid confusion. We define negative or purifying selection as any type of selection where new mutations are selected against, and positive selection as any type of selection where new mutations are advantageous. Diversifying selection is a case of positive selection that will increase variability and is a key mechanism for species divergence; identifying proteins or specific residues experiencing diversifying selection may be important for understanding gene function.

To understand the selective pressures that have shaped genetic variation, protein-coding genes are used to evaluate the relative rates of synonymous and nonsynonymous substitution (Nei & Gojobori, 1986; Miyata & Yasunaga, 1980; Li et al., 1985a). A common approach to this problem is estimate the rates of nonsynonymous (dN or β) and synonymous (dS or α) substitutions. Nonsynonymous mutations can affect protein function and influence the fitness of an organism; while synonymous mutations leave the amino acid sequence unchanged. Under negative or purifying selection, less 'fit' nonsynonymous substitutions accumulate more slowly than synonymous substitutions, and under diversifying or positive selection, the converse is true. Therefore, an important concept in the analysis of coding sequences is that the comparison of relative rates of nonsynonymous and synonymous substitutions can provide information on the type of selection that has acted on a given set of protein-coding sequences. The ratio $\omega = dN/dS$ (also referred to as β/α or K_a/K_s) has become a standard measure of selective pressure; when $\omega \approx 1$ signifies neutral evolution, $\omega < 1$ is negative selection and $\omega > 1$ positive selection.

The estimates of dN significantly different from dS provide convincing evidence for non neutral evolution. An advantage of this approach is that avoid assumptions regarding the demographic history of the population, unlike many "neutrality tests" (Tajima, 1989, 1996; Fu & Li, 1993a,b; Deng & Fu, 1996; Fu, 1997; Misawa & Tajima, 1997; Fay & Wu, 2000), which compare estimates of effective population size obtained using different measures of genetic variation. The field of population genetics in the microbial world (Joyce et al., 2002; Mes, 2008) is still a young area of research. Although the effective population size(N_e) of the ammonia oxidizers could be at a similar range of the 10^9 estimated in *E. coli* (Nei & Graur, 1984), we have to be aware to have enough genetic diversity if we want to use the methods to estimate the dN/dS.

Initial studies of selection pressure relied upon the average dN/dS ratio for the region of interest at the DNA sequence level, either using distance-based methods (Li et al., 1985a; Nei & Gojobori, 1986; Li, 1993; Pamilo & Bianchi, 1993; Comeron, 1995; Yang & Nielsen, 2000) or maximum likelihood methods (Goldman & Yang, 1994; Muse & Gaut, 1994).

In some cases, we are most interested to analyze the selective pressures restricted to a small number of sites. To detect site-specific selection pressures we can rely in *counting methods*, *random effect likelihood models* (REL) or *fixed effect likelihood models* (FEL). Next, a short introduction is provided for each model, although a good review can be found in Pond & Frost (2005):

Counting methods estimates the number of nonsynonymous and synonymous changes that have occurred at each codon throughout the evolutionary history of the sample. This approach was first proposed by Suzuki & Gojobori (1999) and involves reconstructing the ancestral sequences, for example, using parsimony or likelihood-based methods; the latter can take into account the uncertainty in the ancestral reconstructions.

Random effect likelihood models originally described by Nielsen and Yang (1998), involves fitting a distribution of substitution rates across sites and then inferring the rate at which individual sites evolve. When this site-by-site inference is based on the maximum likelihood estimates of the rate parameters, this inference is known as empirical Bayes (Nielsen & Yang, 1998; Yang et al., 2000), whereas when rate class assignments are based on the posterior distribution of rate parameters, this is known as a hierarchical Bayes approach (Huelsenbeck & Dyer, 2004); the latter acknowledges that the rate distribution parameters are sub-ject to error, whereas an empirical Bayes approach treats these parameters as known.

Fixed effect likelihood models are based on fitting substitution rates on a site-by-site basis. Such models are known as fixed effects in the statistical literature. These models can be considered as an extension of the model proposed by Yang & Swanson (2002), who considered two classes of sites specified a priori evolving under different dN/dS. Suzuki & Nei (2004) proposed a model in which the ratio of nonsynonymous to synonymous substitution rates was estimated for each codon using maximum likelihood and a likelihood ratio test (on one degree of freedom) was used to test whether this ratio was significantly different from 1 (neutral).

However, detecting positive selection is generally difficult because positive selection often acts on a few sites and in a short period of evolutionary time, and the signal may be swamped by the ubiquitous negative selection. Another caveat of those methods, is the inability to distinguish in which of the branches selection has occurred, the use of these models is generally geared to the detection of the actual sites that are functionally important or have experienced positive selection with a protein.

For solving this problem a set of methods, referred as *branch-site* tests (Yang & Nielsen, 2002; Pond et al., 2011), offered a model-based phylogenetic hypothesis testing framework for deciding whether or not a lineage (or lineages) of interest had undergone adaptive change. recently Murrell et al. (2012) developed the mixed effects model of evolution (MEME) to detect episodic diversifying selection at individual sites.

When analyzing lineages, the branch-site tests measure selective pressure by ω, the ratio of non-synonymous (dN) to synonymous (dS) substitution rates, and if a proportion of sites in the sequence provides statistically significant support for $\omega > 1$ along the lineages of interest, then episodic positive selection is inferred.
Unfortunately, the original formulation of the method suffered from high rates of false positives when the model assumptions were violated (Zhang, 2004), because the model could misidentify relaxed selective constraints as evidence of diversifying selection, and was subsequently revised to address that shortcoming (Zhang et al., 2005). One of the problems was that the lineages to be tested (*foreground* lineages) were specified a priori, until a recent extension outlined and benchmarked a sequential testing approach to examine whether any single lineage was under selection (Anisimova & Yang, 2007).

Pond et al. (2011) introduced a new set of models in which substitution rates may vary from branch to branch and from site to site. The variation is incorporated via random effects - unobserved strengths of selection at sites and branches are incorporated using a discrete or a discretized parametric probability distribution. Parameters defining the distribution are estimated jointly from all sites using maximum likelihood. Just like existing branch-site methods (Anisimova & Yang, 2007), they use sequential likelihood ratio testing to identify which branches support a model with episodic diversifying selection. Unlike existing methods, however, this approach is unrestricted and considers every possible site profile, thus avoiding some of the prominent issues posed by model misspecification and further allows ω rates to vary independently from branch to branch and site to site.

Based on the brach-site random effects phylogenetic method developed by Pond et al. (2011), MEME, allows the distribution of ω to vary from site to site (the fixed effect) and also from branch to branch at a site (the random effect). This approach provides a qualitative methodological advance over existing approaches which integrate site-to-site and lineage-to-lineage rate variation. MEME can reliably capture the molecular footprints of both episodic and pervasive positive selection, a task for which current models are not well suited.

Analyzing the substitution rate using the site-wise and branch-site models we can answer some important questions (i) what kind of selection is acting over *amoA* gene; (ii) where did this selection happens in the gene in terms of

site; (iii) where did this selection happen in terms of lineages and (iv) are those selective pressures observed different between different environments.

When looking for signatures of molecular evolution in microbial populations, there are two main concerns that have to be carefully considered. First, dN/dS is inappropriate when comparing either very distantly-related strains (dS saturated with multiple substitutions), or very closely-related strains, within which dN/dS is inflated by segregating nonsynonymous polymorphism (Rocha et al., 2006; Kryazhimskiy & Plotkin, 2008). And second, phylogenetic discordance caused by recombination affect likelihood methods for quantifying selection pressure on codon alignments may suffer from high rates of false positives (Anisimova et al., 2003; Shriner et al., 2003). Therefore, before applying any method to estimate the rates in variation in microbial genes, we need to find evidences of recombination and accommodate it in a phylogenetic context splitting the sequence alignments in non-recombinant sequence fragments.

Selective pressures have a big influence on adaptation, and can explain part of the ecological and genetic diversity observed. But, we also have to take in account the gene repertoire of the organisms to fully understand these patterns. Although nowadays is cheap and easy to sequence microbial genomes to perform comparative analyses, still there is the handicap to obtain pure cultures. Luckily, the emerging of metagenomics provided a solution to the problem and we can perform genome independent analyses.

2.6 The use of metagenomics to explore archaeal ammonia oxidation

Metagenomics can be defined as the analysis of genomic DNA from a whole community unlike genomics, which is focused on the analysis of genomic DNA from an individual organism or cell. The term metagenomic was first used in Handelsman et al. (1998), a study of soil microbes using random cloning of environmental DNA. Since then, metagenomics has included any study whereby a whole community is analyzed, e.g., directed studies of 16S rDNA diversity from an environment to isolation and analysis of total DNA from environmental samples without prior cultivation (Chen & Pachter, 2005). Two of the key discoveries related to metagenomics studies have been the ubiquity of proteorhodopsin (Fuhrman et al., 2008) and the discovery of the importance of archaeal ammonia oxidizers (Prosser & Nicol, 2008).

Every metagenomic study can be divided in two different stages (Figure 2.5). First, environmental gene surveys and random shotgun sequencing of all genes at once. The advent of next generation sequencing (NGS) has trig-

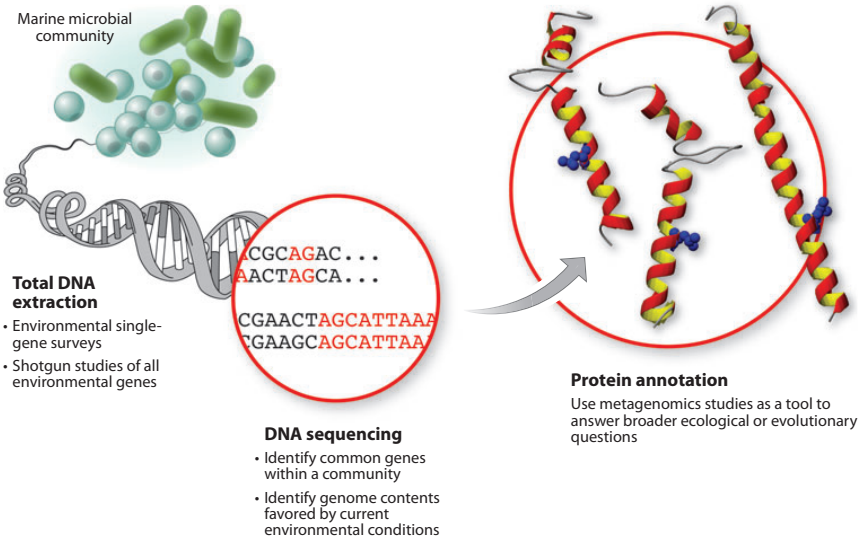

Figure 2.5: Usual metagenomics workflow. Figure from Gilbert & Dupont (2011)

gered a revolution in metagenomic sequencing and analysis. The increased throughput and decrease in costs of sequencing, coupled with additional technological advances have transformed the landscape of metagenomics (Scholz et al., 2012). And second, a bioinformatic analysis to annotate the genes and determine their relationship to the environment. With the annotations we can derive the metabolic state of a community by examining the potential of proteins to consume or produce metabolites. New bioinformatic approaches are needed to handle and analyze all the huge amount of sequence data generated by the NGS techniques.

One of the first metagenomic studies that released their data publicly was the Global Ocean Sampling (GOS) expedition. Venter and colleagues did a sampling from the northwest Atlantic to the eastern tropical Pacific, producing 6.3 billion base pairs from 7.7 million sequence-reads (Rusch et al., 2007). A metagenomic study allows us to cover a broader extension in terms of spatial and temporal scales of variation in marine microbial communities, i.e., a GOS sample represents approximately a period of one week in temporal space, and in geo-spatial space, a few kilometers horizontally and a few meters vertically (Fuhrman, 2009). And the most important, the public release of the sequence data, has facilitated an impressive number of complementary bioinformatic studies by other researchers.

With this large amount of data we can perform a more holistic approach

to the study of marine ecosystems (Karl, 2007). Since a few years ago, network based approaches (Jiang et al., 2012; Faust & Raes, 2012) have started to be used to extract valuable information from the co-ocurrence of rRNAs or individual protein domains involved in biological processes in metagenomic complex systems. The combination of metagenomics and and network based approaches offers a new window to find the missing pieces in biological processes such as the ammonia oxidation carried out by AOA. In this PhD thesis we developed and applied a new approach based on Graphical Gaussian Models to explore functional genes interactions and to confront one of the main challenges of metagenomics, the exploration of the unknown fraction of the metagenomic protein universe.

3
Objectives

The main focus of this PhD work was to explore the ecology and evolution of microbial nitrifiers (ammonia oxidizers). Ammonia oxidation, the first and the rate-limiting step in nitrification, is one of the cornerstones of the cycle. Members from the bacterial and archaeal domains are key players in ammonia oxidation in many different environments. Usually these organisms are found coexisting but the most recent studies suggests that archaeal ammonia oxidizers show an incredible ability to adapt and oxidize ammonia under different environmental conditions and have displaced their bacterial counterparts in terms of importance in the global biogeochemical cycle, providing an avalanche of AOA molecular data (16S rDNA and amoA gene sequences) from very diverse environments worldwide. As far as we don't have enough genomic data to perform an holistic approach using population genomics and reverse ecology to unveil the ecological and evolutionary mechanisms driving the adaptation; we focused our experiments on the amoA gene sequence. Because ammonia monooxygenase is supposed to be the key enzyme in the ammonia oxidation, we applied a combination of community ecology and molecular evolution methods to understand the mechanisms of the diversification patterns observed in the amoA gene. Another unsolved question in the archaeal ammonia oxidation is the unusual biochemistry found in the genome sequences from cultured archaeal ammonia oxidizers. In archaea, all the elements of the bacterial ammonia oxidizing pathway are missing but the genes coding for the presumptive AMO. To unveil missing pathways in this process, we have developed a powerful approach based on graphical models to capture all the functional associations present in metagenomes based in their ecological co-ocurrence.

The detailed objectives and the structure of this PhD dissertation are given

below:

- **Chapter 4** applies community ecology methods to find out the ecological differentiation along the evolutionary history and the temporal diversification patterns of bacterial and archaeal ammonia oxidizers. Our analytical approach identified differences and unveiled the historical patterns of diversification (Fernàndez-Guerra & Casamayor, 2012).

- **Chapter 5** analyzes in detail the selective pressures and adaptive evolution of the main archaeal ammonia oxidizing clusters applying codon-based models at site and lineage level. We identified adaptive differences between AOA from soil and marine environments and generated an hypothesis for the generation and maintenance of an evolutionary AOA *seed bank* driving the adaptive radiation (Manuscript in preparation).

- **Chapter 6** describes T-RFPred, a new software we developed to identify and assign taxonomic information to chromatogram peaks of a T-RFLP genetic fingerprint method for a more comprehensive description of microbial communities. The program helped to describe the AOA communities found in marine metagenomes and can be a valuable tool to carry out targeted metagenomics (Fernàndez-Guerra et al., 2010).

- **Chapter 7** analyzes the ecology and evolution of the phylogenetic partitioning observed in marine AOA. We identify the molecular evidences involved in the photoinhibition and piezophilic adaptations that drives the phylogenetic and ecological partitioning observed along the marine water column (Manuscript in preparation).

- **Chapter 8** describes a new approach to analyze the metagenomic protein universe to unveil unknown functional associations using graphical models. We use the Global Ocean Sampling expedition metagenomes (Rusch et al., 2007) as test dataset to explore the archaeal ammonia oxidation in surface water.We recovered meaningful metabolic associations involved in the ammonia oxidation pathway and map them into the *Nitrosopumilus maritimus* genome providing in silico hypothesis to be further validated in laboratory experiments. (Manuscript submitted).

4

Habitat-Associated Phylogenetic Community Patterns of Microbial Ammonia Oxidizers

Resumen

Los microorganismos que median la oxidación del amonio desempeñan un papel fundamental en la conexión entre la fijación biológica de nitrógeno y las pérdidas anaeróbicas de nitrógeno. Las bacterias y arqueas oxidantes del amonio (AOB y AOA, respectivamente) han colonizado ambientes similares alrededor del mundo. La oxidación del amonio es la etapa limitante en la nitrificación, y la amonio monooxigenasa (AMO) es la enzima clave para este proceso. La ecología molecular de la oxidación del amonio ha sido ampliamente explorada mediante estudios de la subunidad A del gen Amo (*amoA*). En este estudio, hemos explorado la ecología de comunidades de los AOB y AOA, analizando 5776 secuencias génicas del *amoA* aisladas de más de 300 lugares diferente, y clasificando los hábitats utilizando ontologías ambientales. En resumen, la riqueza filogenética es mayor en AOA que en AOB, y los sedimentos contienen la mayor riqueza filogenética mientras que el plancton marino la más baja. También se ha observado que los oxidantes del amonio de agua dulce son filogenéticamente más ricos que sus homólogos marinos. Las comunidades de AOA son más diferentes entre sí que los de AOB y se obser-

van linajes monofiléticos para los sedimentos, suelos y plancton marino para AOA pero no para AOB. Los patrones de diversificación muestran una cladogénesis más constante a través del tiempo para los AOB mientras que en AOA se suceden dos eventos de diversificación rápidos separados por un largo episodio no diversificación. Los índices de diversificación (γ estadístico) para la mayoría de los hábitats indican un $\gamma_{AOA} > \gamma_{AOB}$. Los suelos y sedimentos experimentan estallidos de diversificación tempranos mientras que los hábitats generalmente eutróficos y ricos en amonio, como las aguas residuales y lodos, muestran una aceleración en las tasas de diversificación en el presente. En general, este trabajo muestra por primera vez una visión global de la estructura de la filogenia de comunidades de los dos grupos de oxidadores del amonio (AOA y AOB), siguiendo los más estrictos estándares para su análisis. y proporciona un punto de vista ecológico en los caminos diferenciales evolutivos experimentados por la generalización de oxidantes de amoníaco microorganismos. La imagen que se obtiene de la distribución AOB y AOA en los diferentes hábitats proporciona un nuevo punto de vista para comprender la ecofisiología de los oxidantes del amonio en la Tierra.

Abstract [1]

Microorganisms mediating ammonia oxidation play a fundamental role in the connection between biological nitrogen fixation and anaerobic nitrogen losses. Bacteria and Archaea ammonia oxidizers (AOB and AOA, respectively) have colonized similar habitats worldwide. Ammonia oxidation is the rate-limiting step in nitrification, and the ammonia monooxygenase (Amo) is the key enzyme involved. The molecular ecology of this process has been extensively explored by surveying the gene of the subunit A of the Amo (*amoA* gene). In the present study, we explored the phylogenetic community ecology of AOB and AOA, analyzing 5776 *amoA* gene sequences from > 300 isolation sources, and clustering habitats by environmental ontologies. As a whole, phylogenetic richness was larger in AOA than in AOB, and sediments contained the highest phylogenetic richness whereas marine plankton the lowest. We also observed that freshwater ammonia oxidizers were phylogenetically richer than their marine counterparts. AOA communities were more dissimilar to each other than those of AOB, and consistent monophyletic lineages were observed for sediments, soils, and marine plankton in AOA but not in AOB. The diversification patterns showed a more constant cladogenesis through time for AOB whereas AOA apparently experienced two fast diversification events separated by a long steady-state episode. The diversification rate (γ statistic) for most of the habitats indicated $\gamma_{AOA} > \gamma_{AOB}$. Soil and sediment experienced earlier bursts of diversification whereas habitats usually eutrophic and rich in ammonium such as wastewater and sludge showed accelerated diversification rates towards the present. Overall, this work shows for the first time a global picture of the phylogenetic community structure of both AOB and AOA assemblages following the strictest analytical standards, and provides an ecological view on the differential evolutionary paths experienced by widespread ammonia-oxidizing microorganisms. The emerged picture of AOB and AOA distribution in different habitats provides a new view to understand the ecophysiology of ammonia oxidizers on Earth.

4.1 Introduction

Microbial nitrogen transformations modulate the rate of key ecosystem processes, such as primary production and decomposition (Vitousek et al., 2002). Ammonia oxidation, the first step of nitrification, is a biogeochemical process of global importance in natural and artificial ecosystems worldwide. For many years, ecologists remained perplexed with the nitrifying capacity of

[1] See original publication in Fernàndez-Guerra & Casamayor (2012).

many ecosystems where apparently ammonia oxidizers were far below detection limits, particularly under the most oligotrophic conditions (Olson, 1981). Microbial ammonia oxidation was initially considered to be restricted to a few bacteria, specifically within the phylum Proteobacteria, which under laboratory conditions mostly show an affinity threshold for ammonium higher than the concentrations usually found in situ (Bollmann et al., 2002). Metagenomic studies carried out a few years ago in seawater (Venter et al., 2004) and soil (Treusch et al., 2005) showed a different *amoA* gene related to the phylum Thaumarchaeota, and the presence of the *amoA* gene in widespread Archaea and different habitats has been widely detected since then.

In marine environments, for instance, nitrification accounts for about half of the nitrate consumed by growing phytoplankton at the global scale (Yool et al., 2007) and is responsible for the deep ocean nitrate reservoir (Karl, 2007), the largest pool of reactive nitrogen in the biosphere (Beman et al., 2010; Leininger et al., 2006). In soils, the ammonia-oxidizing archaea (AOA) apparently dominate over ammonia-oxidizing bacteria (AOB) (Leininger et al., 2006). However, ammonium concentrations in many soils have increased in recent years as a result of both land-use changes and increases in atmospheric ammonium concentrations (Rockström et al., 2009), and this may influence the microbial ecology of the nitrification process(Verhamme et al., 2011). Finally, nitrification could remove excessive ammonium nitrogen and prevent lakes from eutrophication (Hagopian & Riley, 1998), and ammonia oxidizers adapted to life in sludge and bioreactors can efficiently help to remove excess of nitrogen (Van Loosdrecht & Jetten, 1998; Mussmann et al., 2011).

Ammonia oxidizers play a fundamental role in the connection between biological N fixation and anaerobic N losses, and are widely detected in a large variety of aquatic and terrestrial environments (Nicol & Schleper, 2006; Auguet et al., 2010). Both AOB and AOA have colonized similar worldwide distributed environments but with different degrees of success in abundance, activity, and distribution(Verhamme et al., 2011), (Prosser & Nicol, 2008; Martens-Habbena et al., 2009; Auguet et al., 2011; Sauder et al., 2011). The molecular ecology of the ammonia oxidation process has been extensively explored surveying the subunit A of the Amo (Auguet et al., 2011; Rotthauwe et al., 1997; Agogué et al., 2008; Francis et al., 2005, 2007). Both AOA and AOB contain the *amoA* gene encoding the alpha subunit of the Amo; however, the gene sequence has evolved separately in each of the phylogenetically distinct but physiologically related ammonia oxidizers microorganisms (Treusch et al., 2005). In the present study, we explored the phylogenetic community ecology of AOB and AOA assemblages, analyzing differences in composition and richness among environments, detecting habitat-phylogeny associations, and unveiling the historical context of the evolutionary events (cladogenesis)

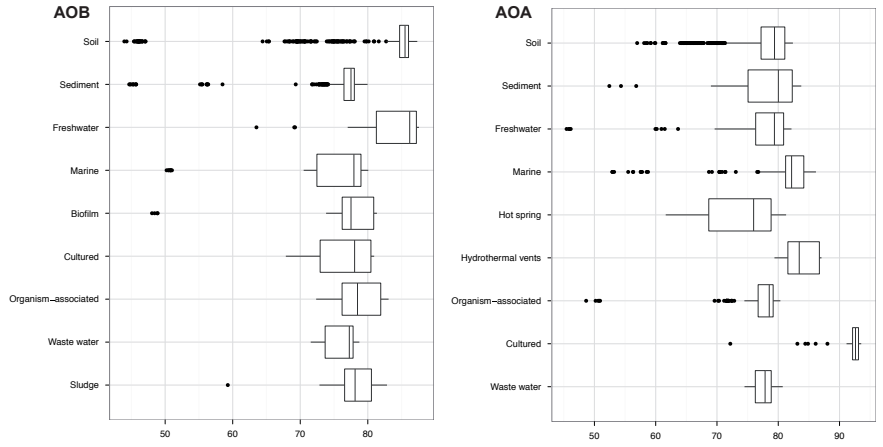

Figure 4.1: Level of *amoA* gene sequence divergence detected in each habitat for bacterial (left panel) and archaeal (right panel) ammonia oxidizers. The boxplot represents all-against-all pairwise alignment identities for the gene sequences compiled from each habitat and for each domain.

captured in the reconstructed *amoA* gene phylogeny.

4.2 Results

Ammonia oxidizers were detected in 11 different habitats worldwide, including soil, sediment, marine plankton, freshwater plankton, organisms-associated, wastewater, sludge, biofilm, hot spring, hydrothermal vent, and the "cultured" habitat (according to the habitat annotation based on the EnvO-Lite description as the microbial assemblages which develop in bioreactors and biofilters). As mentioned in Methods, we assigned the microbial strains isolated in the laboratory to their original habitats. Bacterial and archaeal ammonia oxidizers shared most of these habitats with a few exceptions: AOB were absent in the database for hot springs and hydrothermal vents, whereas the low number of sludge and biofilms AOA sequences recovered from GenBank did not fit the minimal number of sequences required to be included in our analysis. Recent research in these two habitats (Mussmann et al., 2011; Sauder et al., 2011, 2012) is increasing the number of sequences available for future meta-analyses.

To examine the sequence divergence present in each habitat we carried out an all-against-all pairwise alignment (Fig. 4.1). We noticed two trends in the *amoA* gene dataset. First, a median value between 75 and 85% identity was detected in essentially all the natural habitats explored. Second, a substan-

tial number of low identity sequences (below 60% identity) were present in several of the habitats investigated. We explored these outliers to rule out cross-contamination among habitats and found they were specific sequences from each habitat.

To statistically analyze the phylogenetic richness distribution within each community we calculated the phylogenetic diversity (PD) index from the maximum-likelihood inferred trees after correction for unequal sample size (Table S1). Overall, the PD rarefaction curves consistently showed larger phylogenetic richness in AOA than in AOB for an equivalent sampling effort (Fig. 4.2). Neither AOA nor AOB reached the plateau for the PD accumulation, indicating that the currently known phylogenetic richness of the *amoA* gene is far from being fully discovered. The PD values for the AOA and AOB shared habitats (Fig. 4.3A) showed sediments containing the highest phylogenetic richness and marine plankton the lowest. AOA were phylogenetically richer than AOB in both plant- and animal-associated (organism-associated) habitats, and soil. The opposite was found in bioreactors ("cultured" habitat) and wastewater. Interestingly, freshwater AOA and AOB were phylogenetically more diverse than their marine counterparts. AOA in hydrothermal vents and AOB in biofilms showed the lowest PD (Table S1). For most of the explored habitats, AOA were generally more diverse than AOB. Interestingly, the phylogenetic structure captured by the phylogenetic species variability index (PSV) showed AOB to be phylogenetically more overdispersed than AOA in most of the habitats but wastewater (Fig. 4.3B). A general view on the reconstructed tree topologies showed inconsistent phylogenetic clustering for the AOB recovered from the same type of habitats, in agreement with the picture captured by the PSV index. Conversely, three large phylogenetic clusters were found for AOA, i.e., sediment, soil, and marine plankton.

The habitat-phylogeny associations were numerically analyzed with UniFrac distances (UD). The UniFrac matrices were represented graphically for AOB and AOA (Fig. 4.4) to visualize in a quantitative way the community dissimilarity among different habitats. The AOB soil community showed the weakest connections with the remaining nodes in the UD graph, indicating the most distant phylogenetic relatedness (Fig. 4.4), followed by AOB present in sediments. The remaining habitats showed a closer community structure among the different AOB assemblages. Interestingly, the bacterial *amoA* genes found in wastewater were the closest related to the remaining habitats (the strongest connections with the remaining nodes in Fig. 4.4). Conversely, the AOA communities as a whole were more dissimilar to each other, showing weaker interlinks (Fig. 4.4). For AOA, the strongest relationships were observed between freshwater and hot springs assemblages, as well as between wastewater and bioreactors. Soil, sediment and marine plankton were more

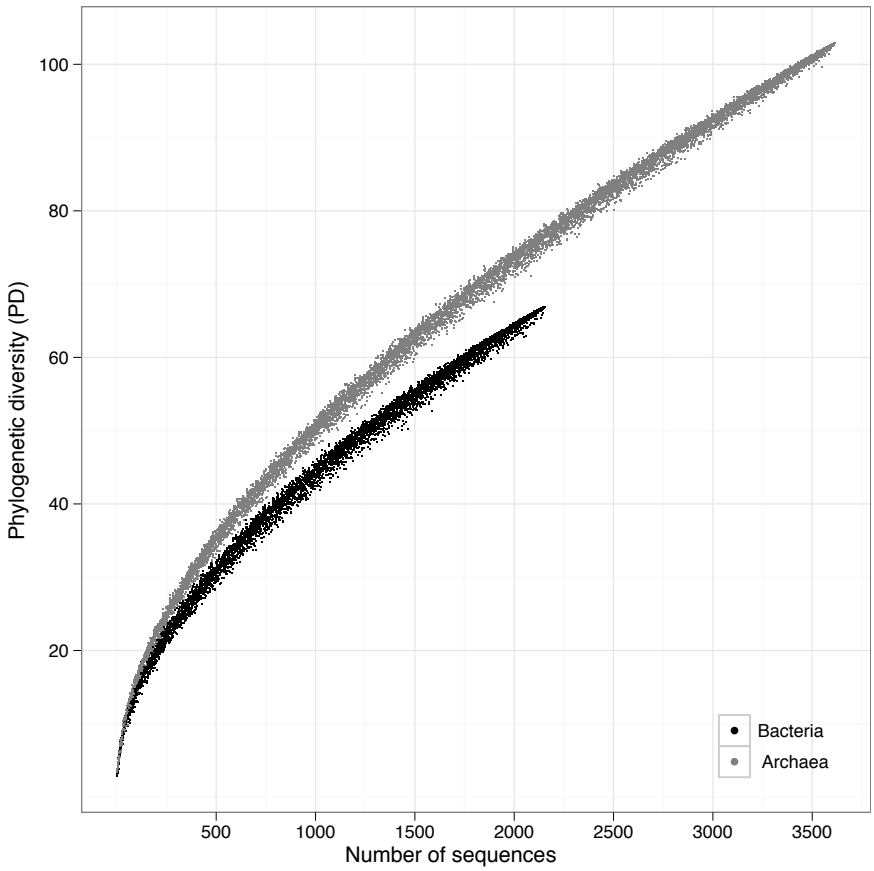

Figure 4.2: Phylogenetic diversity rarefaction curves for the whole dataset of bacterial and archaeal *amoA* gene sequences. Larger phylogenetic richness is observed in AOA than in AOB for an equivalent sampling effort.

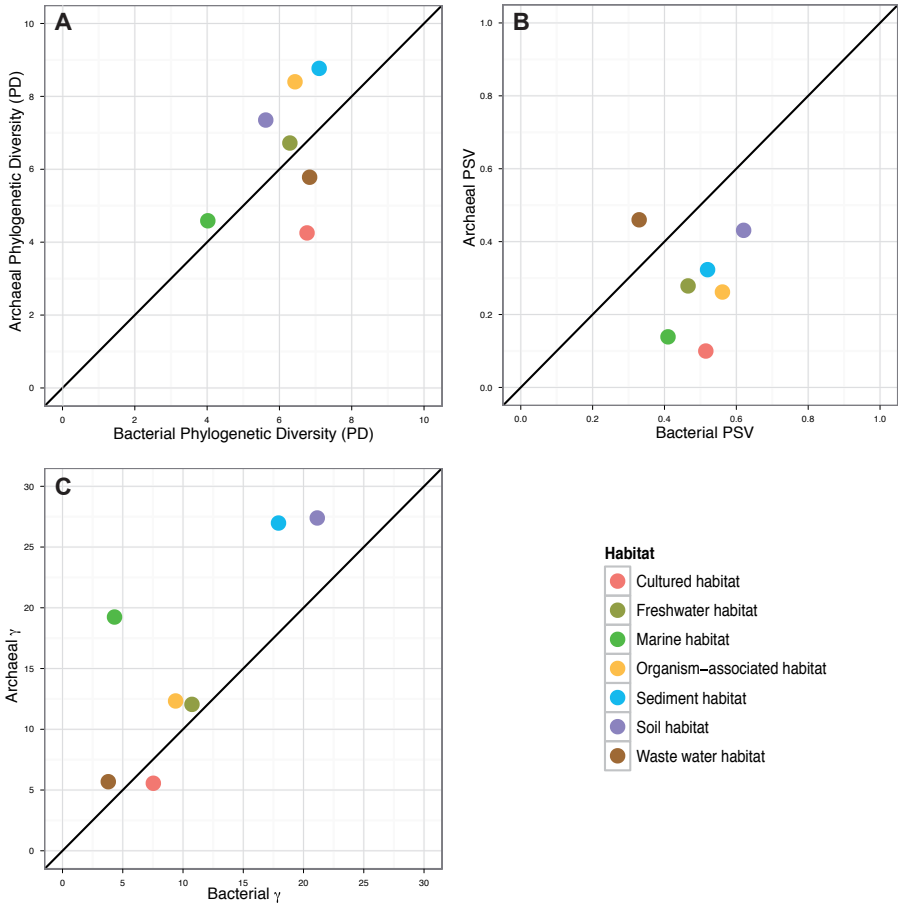

Figure 4.3: Scatter-plot comparing the AOB vs. AOA. A) phylogenetic diversity (PD), B) phylogenetic species variability (PSV) and C) diversification rates (γ-statistic) for the different shared habitats. See Table S1 for data.

unrelated to each other and to the remaining assemblages. Overall, AOB were phylogenetically more closely related among habitats than the archaeal counterparts. In turn, AOA were more phylogenetically clustered by habitat than bacteria.

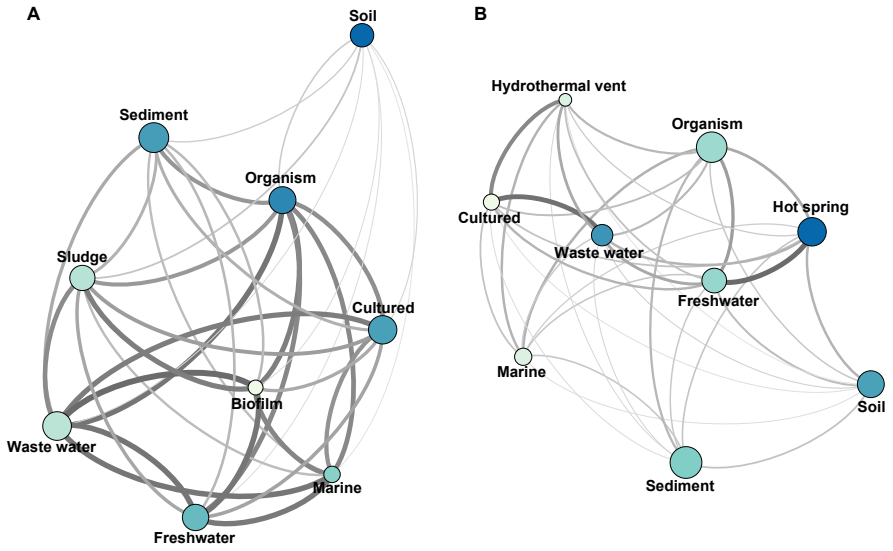

Figure 4.4: Graphical representation of AOB and AOA Unifrac distance matrices. A thicker line represents phylogenetically more similar communities. Phylogenetic diversity (PD) represented as node size, large nodes have larger PD values. Phylogenetic species variability (PSV) represented as node color, darker nodes have larger PSV values. The length of the edges does not contain information.

Finally, we explored the historical context of the evolutionary and diversification processes captured in the reconstructed phylogenies. Soil ammonia oxidizers were placed on the basal position near the phylogenetic tree root both for AOB and AOA. To capture the information contained along the diversification process we represented cladogenesis events versus relative time using lineage-trough time plots (LTT, Fig. 4.5). A consistent increase in the net diversification rate towards present was observed with accelerated recent diversification events and high γ-statistic values for the two life domains (i.e., γ_{AOB} = 34.96, and γ_{AOA} = 45.37). The diversification dynamics in earlier times showed, however, differences between domains with AOB showing a more constant cladogenesis through time whereas AOA apparently experienced two fast diversification events apparently separated by a long steady-state episode (Fig. 4.5). Interestingly, for most of the habitats γ_{AOA} was higher than γ_{AOB} (Fig. 4.3C) and for all the cases the internal phylogeny nodes were

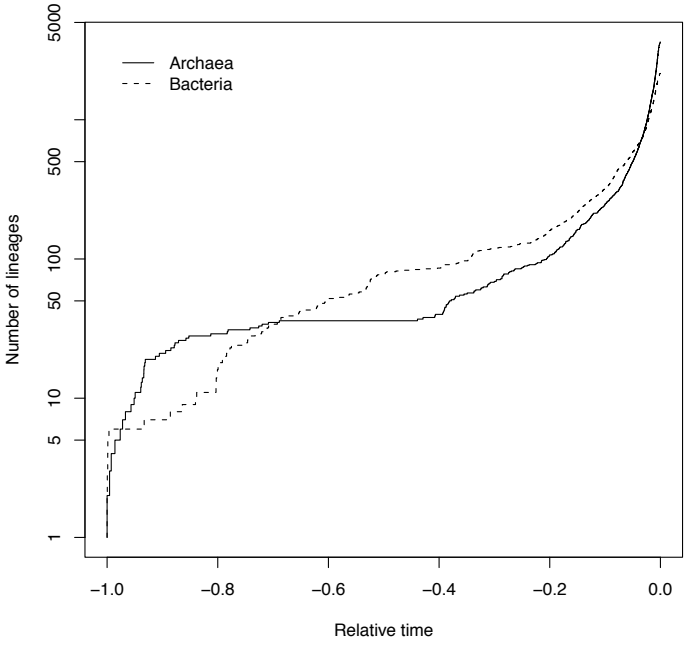

Figure 4.5: Dynamics of the cladogenesis events versus relative time using log-lineage through time plots (LTT) for the inferred phylogenies using the whole dataset of AOA and AOB. Time 0, the present.

closer to the tips than expected under a constant rate of diversification. We explored for each single habitat the historical patterns using a rarefaction analysis to correct for unequal sample size (Fig. 4.6). Interestingly, AOB and AOA in soil and sediment experienced earlier bursts of diversification than in the remaining environments. Afterward, soil AOB experienced a decelerated diversification rate as compared with soil AOA. In addition, AOB in sediments significantly accelerated its diversification events as compared with the soil counterparts. Finally, ammonia oxidizers from habitats usually eutrophic and rich in ammonium such as wastewater and sludge, showed accelerated diversification rates towards the present.

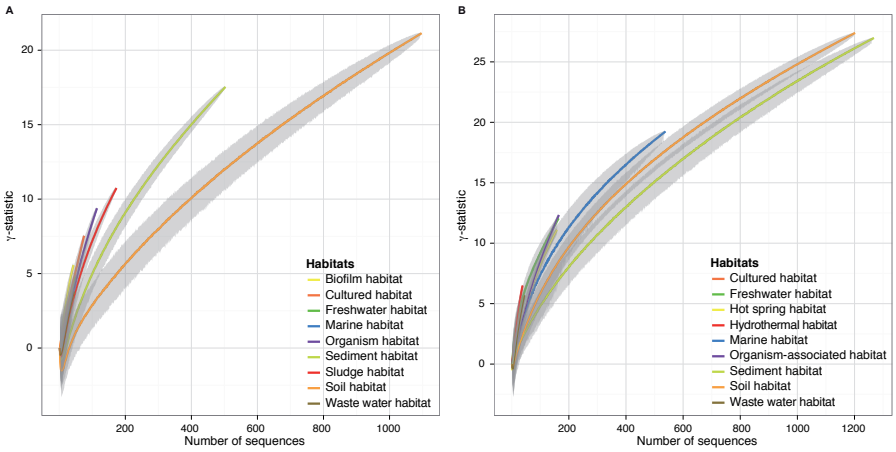

Figure 4.6: Rarefaction curves of the diversification rates (γ-statistic) for AOB and AOA in the different habitats studied.

4.3 Discussion

We have shown consistent differences in the phylogenetic richness among habitats and between domains, with different spatial distribution of the genetic richness (i.e., AOB were phylogenetically more interconnected whereas AOA were more phylogenetically clustered by habitat). These findings suggest differential adaptations of ammonia oxidizers to the large repertory of environmental conditions present in each habitat. In fact, Thaumarchaeota are one of the most widely distributed and abundant groups of microorganism on the planet found in all types of environments, ranging from marine and coastal environments (Agogué et al., 2008; Wuchter et al., 2006), neutral and acid soils (Leininger et al., 2006; Nicol et al., 2008), hot springs (Reigstad et al., 2008; Zhang et al., 2008), remote alpine lakes (Auguet & Casamayor,

2008) and slush layers in ice-covered lakes (Auguet et al., 2012). As recently shown (Pester et al., 2012), AOA have followed different evolution paths suggesting specific physiological adaptations to environments being habitat filtering, salinity and life style (soil and sediment) the main drivers of the community phylogenetic structure as captured by habitat clustering and UniFrac metrics.

Quantitative differences for the community diversity of ammonia oxidizers among different habitats were also important, and especially relevant in the case of soil ammonia oxidizers. Interestingly, the soil AOB community structure was the most dissimilar among the AOB, and for the remaining habitats AOB showed lower diversification rates and higher PSV values. This may be in agreement with the fact that different habitats for different clusters within a single AOB genus have been reported (e.g., *Nitrosospira*, (Prosser & Nicol, 2008) and references therein). AOA, in turn, showed a more heterogeneous community composition among habitats, but specific monophyletic soil groups have been reported (Pester et al., 2012). There are also close links between pH and the relative contributions of bacteria and archaea to soil nitrification, AOA being more favoured at the lowest pH (Nicol et al., 2008). Overall, the accelerated diversification rates in soil AOA may suggest the existence of tight habitat-phylogeny associations in AOA, while in AOB these associations may be not so significant.

AOB and AOA coexist and have to compete for the same resources. In fact, the different degrees of ecological success reported for different assemblages (Prosser & Nicol, 2008; Nicol et al., 2008; Martens-Habbena et al., 2009; Verhamme et al., 2011) suggest that ecological and evolutionary segregation have been acting differently in each domain along the process. AOB for instance, appeared not well adapted to develop in extreme habitats such as hot springs and hydrothermal vents, whereas AOA were apparently less favoured in hypereutrophic sites as sludge (but see, e.g., Mussmann et al. (2011)). A higher substrate affinity and lower tolerance to high substrate concentrations of archaea over bacteria has been detected in one marine isolate (Martens-Habbena et al., 2009) although AOA can survive in high ammonia concentrations in soil (Tourna et al., 2011). Experiments mimicking the conditions of both unfertilized soils and soils receiving moderate and high levels of inorganic fertilizer (Verhamme et al., 2011) showed that ammonium concentration is a more important factor modulating the community structure in AOB than in AOA. The highest ammonium concentrations were also more favorable for the growth of AOB. Overall, nitrogen concentration seems to play a major role in the AOA-AOB interactions. Interestingly, we observed in our analysis that environments permanently rich in nitrogen showed acceleration in the diversification rates that could promote the emergence of new *amoA* variants.

It has been hypothesized that archaea, in general, are better adapted to deal with chronic energetic stress (Valentine, 2007), and this fact may also be captured by the phylogenetic community analysis and inferred population history. Thus, in AOA we observed a LTT plot with a sigmoidal behavior, i.e., two big diversification events, one on the early stage of diversification process and another one on the more recent lineages, and an apparent steady-state between them with constant diversification rate. This feature could be interpreted as an initial high rate diversification process that generated a large number of lineages. These lineages could initially have colonized all available habitats. Then, the diversification remained constant until ecological factors triggered a second diversification event. This second episode could be related to microevolutionary events(Reznick & Ricklefs, 2009) that may facilitate the adaptation of new lineages to new emerged environmental conditions or opportunities. It is interesting to note the high γ-statistic values and the LTT plot shapes in AOB and AOA for soil and sediment, suggesting a competitive race in these habitats. Overall, the low identity percentage at the nucleotide level and the high γ-statistic values suggest that the *amoA* gene is still under an active process of evolution.

The phylogenetic reconstruction was the critical step in the approach. The resulting trees were in agreement with other phylogenetic reconstructions found in the literature showing for instance separated AOB clusters such as *Nitrosomonas*-like and *Nitrosospira*-like, or AOA Cluster-S and Cluster-M (Francis et al., 2005). In fact, the phylogeny of *amoA* genes is largely congruent with the picture derived by 16S rRNA genes analysis, and therefore the habitat-phylogeny distribution patterns found for the *amoA* genes may provide strong hints for the diversity (richness and evenness) of AOB and AOA at the global scale. In addition, the habitat annotation based on the EnvO-Lite ontologies solved two important concerns for massive comparative studies. First, it reduced the number of different ecosystems to a few general habitats that could be easily translated to a globally meaningful classification. Second, it solved the need to be standard, and the ontologies in information sciences are the highest current standard (Hirschman et al., 2008). Being standard means that further studies can be directly and objectively compared, and this will facilitate a comprehensive knowledge base not only for comparative studies but also for integrative database analyses. Finally, although we did an intensive search and a strict filtering process of the publicly available DNA sequences, our results and conclusions are, of course, limited by both the intrinsic biases of the information deposited in databases, and the different methodologies and experimental procedures carried out by different research teams. We cannot rule out, for instance, the inherent biases of PCR amplification and the use of specific *amoA* primer pairs

to generate the data deposited in GenBank from the original environmental samples. To minimize these limitations, the approach taken in this study combined classical phylogeny and community phylogenetics using thousand of sequences from hundred of sites which provided statistically consistent patterns between and within domains at a global scale. These patterns may certainly generate further working hypotheses and help in setting up more accurate experimental designs to improve current knowledge of the ecology and evolution of biological ammonia oxidation.

4.4 Methods

Sequence collection

Two Bioperl scripts were used as wrappers to the Entrez Programming Utilities (Wheeler et al., 2008) to search and retrieve *amoA* sequences from NCBI GenBank database release 178 (June 2010). First, the Esearch utility was used to capture the *amoA* sequences that match the search string *"amo subunit A or ammonia monooxygenase subunit A or ammonia monooxygenase α subunit or Amo α subunit or amoA OR ammonia monooxygenase A or ammonia monooxygenase or ammonium monooxygenase or ammonium monooxygenase A) NOT (genome or chromosome or plasmid)"*. Next, Efetch retrieved the entries found by Esearch in a Genbank formatted flat file to get ancillary environmental information. Overall, 21,603 sequences were retrieved and stored in a postgreSQL database using the associated metadata as information source. Additionally, thaumarcheotal *amoA* sequences recently described from high mountain lakes (Auguet et al., 2012, 2011) and *amoA* from archaeal and bacterial complete sequenced genomes were included in the data set. Sequences were checked to validate the annotation using HMMER 3.0 (Eddy, 2010) in combination with the PFAM 24.0 (Finn et al., 2010) models PF02461 and PF12942 for the bacterial and archaeal *amoA* domains, respectively.

Sequence data preparation

The high quality annotated *amoA* genes dataset was built as follows. Sequences were split by domain, initially resulting in 11,738 sequences for AOA and 9,865 sequences for AOB. Sequences that lack the isolation source tag, CDS tag, those which annotated product was not an ammonia monooxygenase subunit A, and sequences that contained more than 0.1% ambiguous positions were automatically removed. Sequences were further classified by isolation source; all sources with less than five sequences were removed. We ended with 300 different isolation sources (sites), with 153 sites for AOB and

147 sites for AOA. Next, sequences for each site were clustered at 98% identity at the nucleotide level with CD-hit (Li & Godzik, 2006) to reduce redundant sequences. Sequences were clustered by site to keep identical sequences found in different environments. In addition a Perl sequence quality checking script was used to remove sequences considered too short, i.e, those lengths being less than two times the standard deviation of the overall sequence mean length for all sequences in each domain. The final data set contained 3,619 archaeal and 2,157 bacterial *amoA* encoding gene sequences. Final nucleotide lengths were 589±67 for AOA, and 489±88 for AOB.

Envo-Lite annotation

The 300 isolation sources were manually annotated and reduced to 11 different habitat types using environmental ontologies, a standardized project of the of the Genomic Standards Consortium (www.environmentontology.org/). We used the Lite version of EnvO (former Habitat Lite (Hirschman et al., 2008)) that reduces the controlled vocabulary to 20 terms. EnvO provides a controlled and structured vocabulary with defined relationships between its terms allowing efficient and accurate software manipulation, data retrieval and integration. Sequences from laboratory microbial strains were assigned to the original habitat from which they were initially isolated whereas the Envo-lite category "cultured" (i.e., controlled habitat created by humans through laboratory techniques) only contained sequences from both artificial biofilters and bioreactors. We classified separately wastewater and sludge habitats according to the EnvO-lite definitions as follows; wastewater as liquid water that has been adversely affected in quality by anthropogenic influence, and sludge as the residual semi-solid material left from domestic or industrial processes, or wastewater treatment processes. In addition, habitat annotations that only contained a few sequences were combined in a superior hierarchical level (e.g., animal-associated and plant-associated habitats were grouped as organism-associated habitat).

Phylogenetic analysis

The *amoA* sequences were aligned with MAFFT (Katoh et al., 2005), automatically edited with GBlocks (Castresana, 2000) and manually checked and trimmed. A custom Perl script calculated the parameters needed for Gblocks. The final alignment length was 467 positions for AOA and 351 positions for AOB. Substitution saturation in the sequences was checked after plotting distances calculated using Jukes-Cantor, Kimura and raw distances (proportion of different sites), respectively. The plots showed no saturation, so the transition/transversion ratio did not affect the estimated distances.

The *amoA* phylogenetic trees from the nucleotide alignments were inferred by the MPI variant of RaxML v7.2.8 (Stamatakis et al., 2005). Phylogenetic inference was run using the rapid BS algorithm under the GTRCAT model and 20 maximum likelihood searches with 1000 bootstrap replicates to find the best-scoring tree under the GTRGAMMA model. The best phylogenetic tree estimated by RAxML was drawn with iTOL (Letunic & Bork, 2007). Environmental data sets were created and used in iTOL to graphically show the Envo-Lite annotation.

To find the level of sequence identity for each environment, a pairwise alignment all-against-all was carried out for each domain using uclust 1.4 (Edgar, 2010b).

Community phylogenetics

Differences in phylogenetic composition of nitrifying communities among environments were analyzed with UniFrac β-diversity metric (Lozupone & Knight, 2005). To statistically analyze the phylogenetic richness and how diversity was structured in each habitat we calculated phylogenetic diversity (PD) and phylogenetic species variability (PSV) indexes from the inferred trees (Faith, 1992). PD was calculated as the sum of the branch length with 1000 randomizations to avoid the sample size effect. PSV reflects the phylogenetic relationships between taxa, being closer to 1 when all taxa are poorly related (i.e., star phylogeny) and closer to 0 when taxa are closely related. To correct for unequal sample sizes, randomized subsamples for each habitat were run (Barberán & Casamayor, 2010). We also calculated PD rarefaction curves to show how new sequences added larger branch length to the phylogenetic trees.

To graphically show relationships among habitats we used the Gephi 0.8 (Bastian et al., 2009) open source software for graph visualization and analysis and undirected weighted networks on the UniFrac distance matrices and the pairwise alignments. In the UniFrac graph network, vertices correspond to the habitats and the weight of the edges is 1-UD, so the edges represent how similar two communities are.

To estimate divergence time, the original trees were transformed to ultrametric trees through the mean path length method (MPL) (Britton et al., 2007) as rate smoothing technique. We scaled the tree root at relative time 1, and then the tree was calibrated using the root age value. In order to visualize the events of diversification and to measure their changes among habitats we plotted lineage-through-time plots and calculated the γ-statistic (Pybus & Harvey, 2000). For diversification events constant through time, the parameter γ equals zero and a straight line in LTT is expected. If the diversification slowed

then, γ<0 and the LTT plot lays above the straight line (the tree internal nodes are closer to the root than expected under a constant rate of diversification); γ>0 indicates acceleration through time in the rate of lineages accumulation (the nodes are closer to the tips than expected). Rarefaction curves were calculated with the γ value for each habitat.

All analyses were carried out in the R environment (http://www.r-project.org/) using APE (Paradis et al., 2004) and Picante (Kembel et al., 2010) packages.

Acknowledgements

We thank Albert Barberán and JC Auguet for help with data handling, and Jordi Catalan for support. Constructive comments by anonymous reviewers and the editor are acknowledged.

Appendix [2]

[2]See more Supplementary Information in Fernàndez-Guerra & Casamayor (2012).

5
Evolutionary Patterns in Archaeal Ammonia Oxidizers

Resumen

La *amoA* es uno de los genes clave implicados en la oxidación del amonio y presenta una distribución muy diversa y abundante en los diferentes ambientes del planeta. En el siguiente estudio, analizamos los procesos evolutivos moleculares implicados en la alta capacidad diversificadora de este gen; aunque el gen del *amoA* se encuentra bajo los efectos de la selección purificadora, hemos encontrado evidencias de selección episódica diversificadora en codones individuales así como en linajes. Hemos observado eventos de selección postiva diversificadora seguido de periodos de conservación (selección homogeneizadora) como un mecanismo para la generación y mantenimiento de un *seed bank* evolutivo del gen de la *amoA* como un mecanismo para la radiación adaptativa en los diferentes hábitats.

Abstract [1]

The *amoA* is one of the key genes involved in ammonia oxidation and has been found to be very diverse and abundant in different habitats. In the present study, we explored the molecular evolution processes underlying the ecological succesful diversification processes observed in ammonia oxidizing archaea (AOA). Although treated as a whole the *amoA* gene was under purifying selection processes, we found evidences of episodic diversifying selection both on individual codons and in individual phylogenetic lineages. We observed diversification events mediated by bursts of positive selection followed by extensive conservation (homogenizing selection) as a mechanism for generation and maintenance of an *amoA seed bank* as the source for the adaptive radiation found along ecological diverse habitats. We also identified the individual codons that evolved differentially and that were involved in the *amoA* cladogenesis and in the adaptive processes that may determine AOA habitat partitioning.

5.1 Introduction

Ammonia-oxidizing Archaea (AOA) have been widely detected in a large variety of aquatic and terrestrial environments (see a recent metaanalysis in Fernàndez-Guerra & Casamayor (2012)). Recently, different global *amoA* phylogenies have been inferred (Gubry-Rangin et al., 2011; Biller et al., 2012); and a consensus phylogeny and a new nomenclature for the groups observed has been proposed (Pester et al., 2012). The current knowledge splits the archaeal *amoA* gene sequences cluster into five major phylogenetic clades, which tend to cluster sequences from the same habitat Pester et al. (2012). These phylogenies capture the diversification of *amoA* but an explanation for the molecular processes underlying the *amoA* cladogenesis and the habitat adaptation are missed.

Diversification and adaptation are the result of the evolution in protein-coding regions (Hurst et al., 2006) and understanding the patterns behind these evolutionary processes from molecular data, at individual protein sites (Nielsen & Yang, 1998; Yang et al., 2000; Murrell et al., 2012) and over phylogenetic lineages (Yang & Nielsen, 2002; Pond & Frost, 2005; Pond et al., 2011), is an area of active research in genetics studies on viruses and eukaryotes. The application in microbial ecology of this new view to understand how the evolutionary pressures in key functional proteins determine the ecological patterns observed has not been explored so far. A simple and robust way

[1] Fernàndez-Guerra, A, EO Casamayor. **Manuscript in preparation.**

to quantify these molecular evolutionary pressures is the ratio of substitution rates for synonymous sites (dS or α), which are presumed to be neutral, and for non-synonymous sites (dN or β), which possibly experience selection. When natural selection promotes changes in the protein sequence the $dN/dS > 1$ (positive selection) and when natural selection suppresses protein changes the $dN/dS < 1$ (purifying selection). Before using the dN/dS ratio in prokaryotes a few concerns have to be carefully considered, such as the high level of recombination in prokaryotic genomes (Kreuzer, 2005) and the sequences analyzed has to be from divergent lineages (Kryazhimskiy & Plotkin, 2008).

In a recent work Biller et al. (2012), has analyzed the selective pressures of the *amoA* gene, as the mean of the ratio of non-synonymous to synonymous substitutions (dN/dS) by counting methods -Pond 2005b- and found that *amoA* gene is under purifying selection and any site was detected to experience positive selection. However, selective pressures can vary both over sites and time, resulting in localized bursts of selection within a subset of sites and a small number of lineages. In fact, recently it has been suggested that natural selection could be basically episodic (Murrell et al., 2012).

In the present study, we explored the evolutionary patterns of three of the main clusters obtained from the consensus AOA phylogeny. We analyzed the *Nitrosopumilus* cluster (Könneke et al., 2005) as a representative of a marine AOA generally found in open ocean and adapted to live in oligotrophic environments, the *Nitrososphaera* cluster (Tourna et al., 2011) an ammonia oxidizing archaeon from soil and *Nitrosotalea* cluster (Lehtovirta-Morley et al., 2011) an obligate acidophilic ammonia oxidizer from soil. We identify events of episodic diversifying selection both at individual sites and in individual lineages, the mechanisms for generation and maintenance of an evolutionary *amoA seed bank*, and different evolving sites related to cladogenesis and habitat partitioning of the three main AOA clusters.

5.2 Methods

Sequence collection and environmental annotation

The *amoA* gene sequences were obtained from the ARB database distributed by (Pester et al., 2012); in their original work those sequences were used to infer the most up-to-date and rigorous consensus AOA phylogeny. We removed 20 sequences containing ambiguities from the 735 sequences present in the initial data set, and we followed the new nomenclature system for the AOA lineages proposed by the authors. To make the annotation the most accurate possible we environmentally re-annotated the sequences combining (and redefining when necessary) the original annotation with the terms found in Habitat-

lite (Hirschman et al., 2008). When combined the environmental annotation with the main *amoA* clusters we were able to generate several data sets using the consensus tree topology as a reference to split the database. The first data set comprised the three most abundant archaeal *amoA* clusters, ,i.e., *Nitrosopumilus*-like with 352 sequences, *Nitrosotalea*-like with 39 sequences, and *Nitrososphaerea*-like with 310 sequences. The second data set was generated using the environmental annotation, choosing the five habitats with the highest number of sequences., i.e., soil (180 sequences), marine sediment (150), estuarine sediment (148), marine planktonic (95, referred as marine), and hot spring (39 sequences). Finally, the third data set was conformed by cluster specific habitat sequences from *Nitrosopumilus* and *Nitrososphaera*. In *Nitrosopumilus* cluster, 107 sequences were from marine sediment, 90 from marine plankton and 80 from estuarine sediments. For *Nitrososphaera*, 135 sequences were from soil habitats, 41 from marine sediment and 51 from estuarine sediment.

We used the *amoA* genomic sequence from *Nitrosopumilus* (YP_001582834.1) aas reference to naming positions in further analysis.

Sequence alignment and phylogenetic inference

Although the sequences distributed within the ARB database were already aligned, we used our own alignment approach to achieve the high requirements in alignment quality for codon-based analyses. First, we checked if selected sequences were on frame doing a BLAST search of the nucleotide sequences against their respective amino acid sequences using bl2seq with the algorithm tblastn from NCBI Blast+ 2.2.25 package (Camacho et al., 2009). Then, the nucleotide sequences were spliced following the coordinates of the blast results. Later to minimize the codon alignments errors and improve the detection of the selection events (Jordan & Goldman, 2011; Privman et al., 2011) we used an in-house modified version of GUIDANCE (Penn et al., 2010). This customized version of GUIDANCE takes profit of parallelization in some of the steps required by GUIDANCE and uses RAxML as fast maximum-likelihood phylogenetic method to infer the guiding trees for MUSCLE (Edgar, 2010a). The high quality codon alignment after applying the filtering algorithms implemented in GUIDANCE to deal with the alignment uncertainty resulted in 594 positions. We generated consensus sequences for the alignments; aminoacid composition at a site was resolved using the most frequent character at that site for each alignment. Sequence logo (Crooks et al., 2004) for those alignments can be found in Figure S1-S4.

Screening for recombination

We screened all cluster *amoA* alignments using the GARD method implemented in HYPHY (Pond et al., 2005) to find evidences of recombination. Recombination events may mislead selection analysis as each partition could have different rates; to correct this effect, selection codon based analysis had to take in account a tree topology for each partition (Scheffler et al., 2006). We did a GARD screening on 5 different random data sets picking up 20 sequences from each cluster. We repeated GARD analysis at least five times for each random data set to check convergence and test if the recombination breakpoints found were stable due to the aleatory nature of GARD.

HYPHY implemented the Kishino-Hasegawa test (KH) (Kishino & Hasegawa, 1989) to test the congruence of topologies; we inferred separately a ML tree for each partition using RAxML 7.3.0 (Stamatakis et al., 2005) performing 1000 maximum likelihood searches with 1000 bootstrap replicates and with the consensus topology as starting tree to find the best-scoring tree under the GTRGAMMA model for each partition of the alignment determined by the breakpoint location and we calculated the sitewise log likelihoods. To test that the incongruences found by the KH test were caused by a different ratio in each side of the breakpoint, p-values of each alignment combination according to the Approximately Unbiased (AU) test using multiscale bootstrapping (Shimodaira, 2002) and the Shimodaira-Hasegawa (SH) test (Shimodaira & Hasegawa, 1999) were calculated using CONSEL (Shimodaira & Hasegawa, 2001). Tree topologies inferred for each partition were used for later codon based analyses and we calculated the tree length (S) for each partition (Table 5.1)

GARD reported breakpoints at positions 298, 299, 300, 306 and 311 in the five random data sets we screened. Recombination signal detected by GARD suggested a breakpoint nearby position 300. We partitioned *amoA* alignments at position 300 and we inferred a maximum likelihood tree for each partition as previously described. Then, we tested the inferred tree topologies for incongruences using the KH test implemented in HYPHY, in all cases KH test reported a $p < 0.01$ supporting the presence of a breakpoint. In addition, CONSEL reported $p < 0.01$ for AU and SH tests.

Codon based analyses

HyPhy package for testing hypothesis using phylogenies was used to detect signatures of positive and negative selection and differential adaptive evolution from the *amoA* codon alignments estimating the ratio of non-synonymous (dN or β) and synonymous substitution rates (dS or α). First, we performed pairwise comparisons of genetic distance between clusters, habitats and cluser

specific habitat to analyze different populations and not segregating polymorphisms (Kryazhimskiy & Plotkin, 2008). We calculated the F_{ST} (distance based; the distance matrices were calculated using the TN93 genetic distance) pairwise metrics using the methods implemented in HYPHY (Zárate et al., 2007) with 1000 bootstraps and 1000 permutations.

Then we searched for the best nucleotide model to be combined with the codon model MG94 (Muse-Gaut 1994). Nucleotide models that fitted best each data set by the AICc criterion are shown in Table 5.1 (Model column)

To estimate the rates of non-synonymous and synonymous in each site of the alignment to identify which codon sites were under pervasive negative selection we used the Fixed Effects Likelihood for terminal (FEL) and internal (iFEL) branches taking in account recombination. As described in (Pond & Frost, 2005) FEL methods outperforms Random Effects Likelihood (REL) methods and counting methods (SLAC). FEL approach didn't suffer from so many false positive as the other methods and was better for detecting the patterns of rate variation. IFEL was essentially the same as FEL, except that selection was only tested along the internal branches of the phylogeny (Pond et al., 2006).

One of FEL limitation was that the method assumed the same dN/dS ratio to all branches, which meant, that didn't exist a lineage-to-lineage variation in dN/dS. These assumptions affected to the detection of those clades that could be under a different regime of selection but the signal was lost in the background noise of the overall tree. To avoid this effect and find out which sites were under episodic or pervasive positive selection at the level of an individual site we applied the Mixed Effects Model of Evolution (MEME) (Murrell et al., 2012) a mixed effects model of evolution that allows the distribution of ω to vary from site to site and also from branch to branch at a site. To detect which lineages -MEME, only detected individual sites- were under episodic diversifying selection (EDS) we applied the BranchSiteREL method (Pond et al., 2011).

HyPhy provided the possibility to compare differential evolution in different samples -in our scenario, we defined *amoA* clusters, habitats, cluster associated habitats as different samples- following the hypothesis that at a given site, the dN/dS ratio differed between the two samples, on the entire tree and on the internal branches as described in (Pond et al., 2006). This method only detected sites subject to different selective pressures in both samples, regardless of which residue appeared to be selected for.

All codon-based analyses were run on the non-recombinant fragments to take account of recombination separately. We plotted the lineages detected by BranchSiteREL to be under episodic or pervasive positive selection on the consensus *amoA* phylogeny. Trees were graphically represented with iTOL (Letu-

nic & Bork, 2007).

In order to deal with convergence problems, most analyses were run at least twice with different initial conditions, and the run with the best likelihood for each was kept.

Hive plot were used to show the distribution of sites detected under Episodic Diversifying Selection (see details of this type of representation in Appendix A.

5.3 Results

Pairwise comparisons of the genetic distance calculated by the F_{ST} statistic, reported that the all data set included in the analysis diverged significantly from one another at $p < 0.001$. And the tree lengths (S) calculated as the expected substitution per codon site along the tree (Table 5.1) assured medium sequence divergence (Anisimova et al., 2002). Accurate prediction in similar sequences is possible when we include a large number of lineages. For example a $S = 15.8$ in *Nitrosopumilus* partition means an average branch length of $S/(2T-3) \approx 0.02$ nucleotide substitutions per codon (where T are the number of taxa).

Signatures of purifying selection

Overall, the *amoA* alignment data set analyzed in the present work were under the effects of purifying selection (or negative selection), alignments had a mean $dN/dS < 1$ (ranging from 0.032 on the marine sediment alignment from *Nitrososphaera* to the maximum value of 0.078 for the hot springs alignments).

On Table 5.1 are summarized the results after applying FEL methods to internal/terminal branches to analyze the *amoA* clusters (taxonomy), habitats and cluster specific habitats.

On the cluster level, *Nitrosopumilus* had 183/185 negatively selected sites (Table SA1 and SA15), *Nitrosotalea* had 103/124 (Table SA2 and SA16) and *Nitrososphaera* 185/183 (Table SA3 and SA17) at $p \leq 0.05$.

When habitat alignments were analyzed using FEL methods, the number of sites under purifying selection for internal/terminal branches for marine habitat was 175/179 (Table SA4 and SA18), 182/186 for marine sediments (Table SA5 and SA19), 180/184 for estuarine sediments (Table SA6 and SA20), 144/158 for hot springs (Table SA7 and SA21) and 184/186 for soils (Table SA8 and SA22).

In cluster specific habitats for *Nitrosopumilus*, iFEL/FEL reported 171/175 negatively selected sites on marine habitats (Table SA20 and SA23), 164/171

Table 5.1: Sequences analyzed in this study, classified by *amoA* cluster, habitat and cluster specific habitat. N=number of sequences; S=Tree length as expected expected substitution per codon site; Model: Best nucleotide model by AIC_c; MEME, FEL, iFEL, dN internal, dN Tips Only: number of codons detected at $p \leq 0.05$

Data set	N	S	Model	Mean dN/dS	MEME	FEL	iFEL	dN Intenal	dN Tips Only
amoA Cluster									
N'pumilus	352	15.8/16.8	012310	0.044	25	185	183	64	92
N'sphaera	310	10.3/17.2	012343	0.045	16	183	185	45	105
N'talea	37	1.6/2.4	012313	0.067	6	124	103	39	29
Habitat									
Marine	95	4.8/5.0	012310	0.043	8	179	175	58	40
Marine sediment	150	8.6/10.1	012314	0.047	19	186	182	77	48
Estuarine sediment 138	8.1/10.4	012313	0.045	20	184	180	74	47	
Hot spring	39	2.70/3.53	010212	0.078	9	158	144	69	40
Soil	180	8.49/12.0	012313	0.044	11	186	184	62	73
amoA cluster specific habitat									
N'pumilus marine	90	4.2/4.1	012310	0.040	8	175	171	39	46
N'pumilus marine sediment	107	5.4/5.5	012310	0.053	16	171	164	51	45
N'pumilus estuarine sediment	80	4.2/4.8	012010	0.048	11	172	154	43	37
N'sphaera soil	168	7.5/10.8	012313	0.039	9	184	182	38	83
N'sphaera marine sediment	41	2.8/4.0	012313	0.032	2	179	169	34	27
N'sphaera estuarine sediment	51	3.3/4.7	012313	0.036	5	179	175	33	30

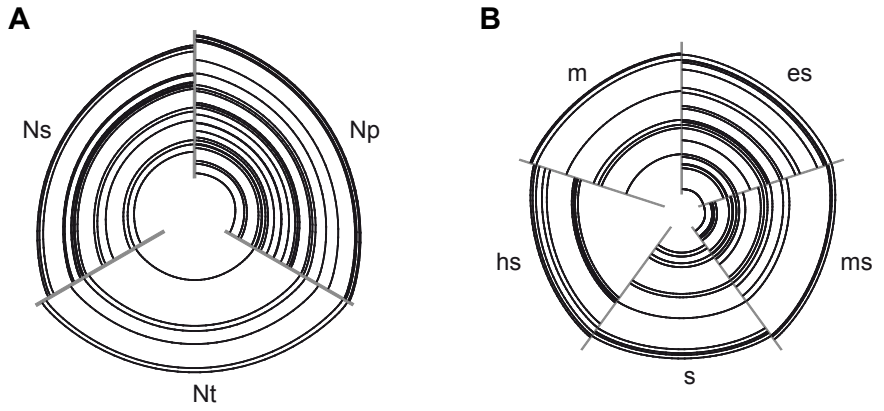

Figure 5.1: Hive plot with the distribution of sites detected, $p \leq 0.05$, under Episodic Diversifying Selection on *amoA* clusters (A) and habitats (B). Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 from the genomic sequence of *Nitrosopumilus*; codon 1 located at the axis inner part of the hive plot. Ns: *Nitrosopumilus*; Ns=*Nitrososphaera*; Nt=*Nitrosotalea*; m=marine; ms=marine sediment; es=estuarine sediment; hs=hot spring; s=soil.

on marine sediment (Table SA21 and SA24) and 154/172 on estuarine sediments (Table SA22 and SA25); in *Nitrososphaera* iFEL/FEL reported 182/184 negatively selected sites on soil habitats (Table SA12 and SA26), 169/179 on marine sediment (Table SA13 and SA27) and 175/179 on estuarine sediments (Table SA14 and SA28).

IFEL allows the estimation of dN and dS separately in internal and terminal branches of the tree, this allowed us to identify if recent non-synonymous substitutions (terminal branches) were not represented on internal branches or the other way around (Table 5.1). In five alignments, *Nitrosopumilus* (92:64), *Nitrososophaera* (105:45), soil habitat (73:62), *Nitrosopumilus* from marine habitat (46:39) and *Nitrososphaera* from soil habitat (83:38) had more codons with only recent non-synonymous substitutions than codons with substitutions on internal branches. Neither FEL nor iFEL detected any site under pervasive positive selection.

Signatures of Episodic Diversifying Selection

MEME detected signals of episodic diversifying selection ($p \leq 0.05$) on 25 sites for *Nitrosopumilus* (Table SB1), 6 for *Nitrosotalea* (Table SB2) and 16 for *Nitrososphaera* (Table SB3). Figure 5.1A shows a hive plot (Krzywinski et al., 2012) with the distribution of the sites under positive selection detected by MEME in each *amoA* cluster (Figure S5 shows detailed results on the proportion of

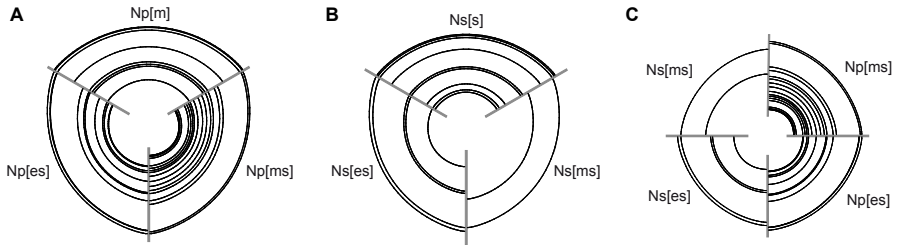

Figure 5.2: Hive plot with the distribution of sites detected, $p \leq 0.05$, under Episodic Diversifying Selection on *amoA* clusters specific habitats (A, B) and clusters only including sediment related habitats (C). Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 form the genomic sequence of *Nitrosopumilus*; codon 1 located at the axis inner part of the hive plot. Np: *Nitrosopumilus*; Np=*Nitrososphaera*; Nt=*Nitrosotalea*; m=marine; ms=marine sediment; es=estuarine sediment; s=soil.

branches under EDS).

We found codon {47, 131} to be under diversifying selection in all three taxa. Codon {47} strength of diversifying selection (β^+/α) is 19.68 and 0.7% of branches (q^+) were under diversifying selection (Table S1). In *Nitrososphaera* had a $\beta^+/\alpha = 6.61$ and $q^+ = 4.8\%$ and *Nitrosotalea* had a $\beta^+/\alpha = 71.14$ and $q^+ = 2\%$ at this site. For codon {131}, *Nitrosopumilus* had a $\beta^+/\alpha = 78.59$ and a $q^+ = 1.8\%$ of branches were under diversifying selection. *Nitrososphaera* had a $\beta^+/\alpha = 34.84$ and $q^+ = 1.2\%$ and *Nitrosotalea* had a $\beta^+/\alpha = 106.31$ and $q^+ = 7.2\%$ at this site.

While some of the codons identified to be under diversifying selection were common within clusters, others were cluster specific. Codons {7, 20, 24, 42, 48, 57, 66, 77, 78, 89, 103, 106, 126, 166, 194, 195, 201} were only detected on *Nitrospumilus*; while codons {37, 81, 94, 125, 129, 134, 135, 145, 178, 184, 187} were positively identified only in *Nitrosotalea*; and codon {152} only on *Nitrosotalea*.

When MEME was applied to the habitat alignments (hive plot on Figure 5.2B and detailed results on Figure S6) reported 8 sites for marine (Table SB4), 19 for marine sediment (Table SB5), 20 for estuarine sediment (Table SB6), 9 for hot spring (Table SB7) and 11 for soil (Table SB8). Codon {192} was detected by MEME for all habitats. This codon had a $\beta^+/\alpha = 30.25$ and $q^+ = 2\%$ in marine habitats; $\beta^+/\alpha = 57.38$ and $q^+ = 0.8\%$ in marine sediment; $\beta^+/\alpha = 1343.93$ and $q^+ = 2.17\%$ in estuarine sediment; $\beta^+/\alpha = 48.32$ and $q^+ = 3.26\%$ in hot spring; and $\beta^+/\alpha = 43.85$ and $q^+ = 1.1\%$ in soil.

Codons detected under diversifying selection specific by habitat were {147, 201} for marine habitat, codons {17, 20, 24, 48, 66, 89, 194} for marine sediment, codons {43, 77, 103, 126, 145, 152, 188, 191} for estuarine sediment, codons {125,

129, 134, 171, 195} for hot spring and codons {31, 184} for soil.

On cluster specific habitats, Figure 5.3 and Figure S7 for detailed results, MEME reported 8 sites in marine habitats (Table SB9), 16 in marine sediment (Table SB10) and 11 in estuarine sediments (Table SB11) for *Nitrosopumilus* cluster; while for *Nitrososphaera*, MEME reported 9 sites in soils (Table SB12), 2 on marine sediments (Table SB13) and 5 on estuarine sediments (Table SB14). Codons {57,199} were detected by MEME to be under diversifying selection on all *Nitrospumilus* habitats; with a $\beta^+/\alpha = 193.74$ and $q^+ = 0.7\%$ in marine, $\beta^+/\alpha = 51.05$ and $q^+ = 8.8\%$ in marine sediment, $\beta^+/\alpha = 312.28$ and $q^+ = 9.14\%$ in estuarine sediment for codon {57}. For codon {199}, MEME reported a $\beta^+/\alpha = 71.06$ and $q^+ = 0.7\%$ in marine, $\beta^+/\alpha = 93.22$ and $q^+ = 1.6\%$ in marine sediment, $\beta^+/\alpha = 21.74$ and $q^+ = 2.5\%$ in estuarine sediment.

Codon 178 was detected by MEME on all *Nitrososphaera* habitats and reported a $\beta^+/\alpha = 67.057$ and $q^+ = 0.8\%$ in soil, $\beta^+/\alpha = 94.59$ and $q^+ = 1.4\%$ in marine sediment, $\beta^+/\alpha = 973.30$ and $q^+ = 1.07\%$ in estuarine sediment. For *Nitrosopumilus* specific habitats, codons only detected under diversifying selection in each habitat were {94, 147, 201} in marine habitat, codons {17, 20, 24, 48, 66, 78, 89, 131, 194} in marine sediment habitats and codons {43, 77, 103, 191} in estuarine sediments. For *Nitrososphaera* habitats, codons {31, 53, 147, 184} in soils and codon {112} in marine sediment were only positively detected on those habitats.

Lineages under episodic diversifying selection

For every main cluster in the consensus *amoA* phylogeny (Figure 5.3), Branch-SiteREL method found evidences of lineages under episodic diversifying selection with a corrected $p \leq 0.05$. Detailed results for the branch-level mixture of negative, (nearly) neutral and positive selection models values could be found in Tables SD1-3.

The analysis reported for *Nitrospumilus* evidences of 20 lineages where the strength of diversifying selection (ω^+) > 40 and the proportion of sites under selection (q^+) range varies from 0.5 to 10.90%. Five of the lineages detected were members of the *subcluster 9*. All other lineages detected are spread around the other subclusters, e.g. branch-site methods detected two lineages for *subclusters 1*, *2*, and *3*. Some of the lineages with higher q^+ were Node612 with a $q^+ = 10.90\%$ and $\omega^+ = 154.577$, an ancestral lineage from *Nitrosopumilus subcluster 9.1*; Node568 with $q^+ = 4.1\%$ and $\omega^+ = 3333.960$ and Node399 with $q^+ = 4.5\%$ and $\omega^+ = 1999.220$.

In *Nitrososphaera* cluster, 4 branches were detected to be under EDS. All 4 branches had a *omega* > 200; Node28 had a $q^+ = 1.1\%$ and $\omega^+ = 10000$;

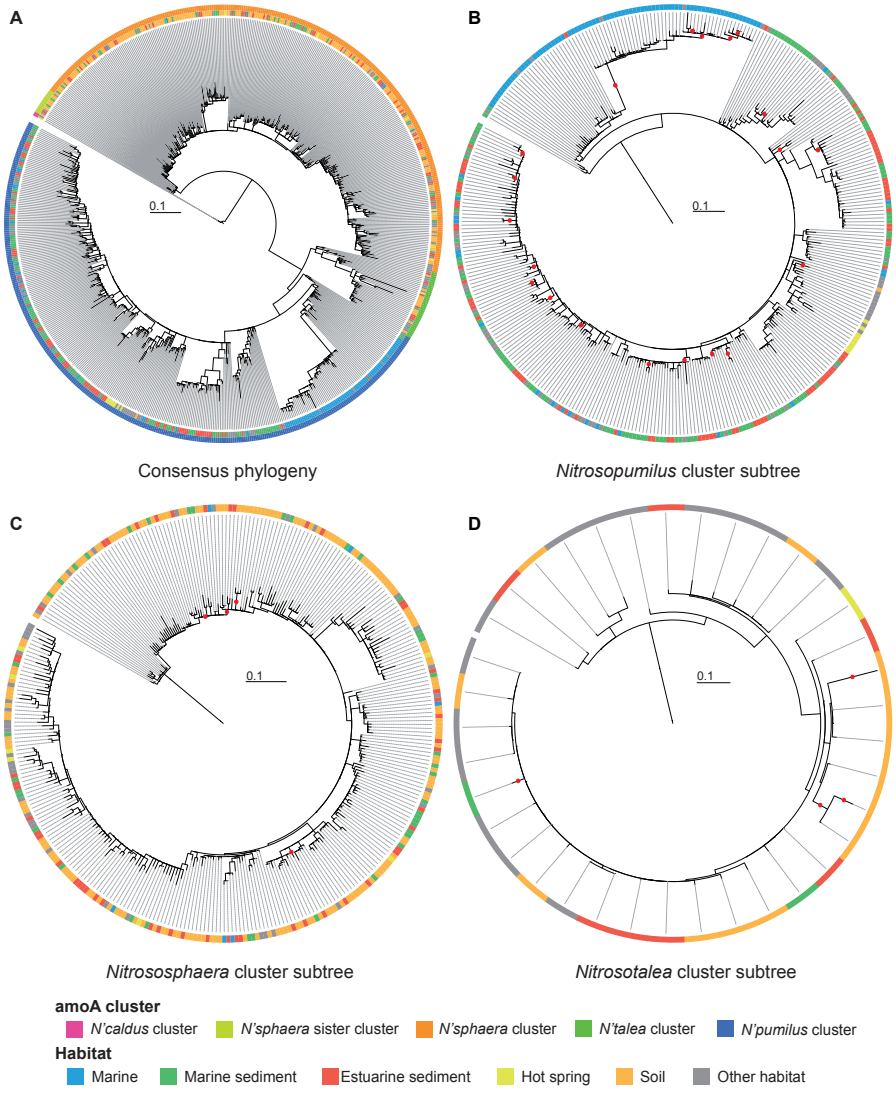

Figure 5.3: Consensus phylogeny from Pester et al. (2012) (A). Cluster subtrees from consensus phylogeny for *Nitrosopumilus* (B), *Nitrososphaera* (C) and *Nitrosotalea* (D). Branches detected under episodic diversifying selection by the sequential test at $p \leq 0.05$ are marked with a red circle. Each tree is scaled on the expected substitution per codon site

Node34 had a $q^+ = 0.5$ and a $\omega^+ = 10000$; Node299 had a $q^+ = 1.8\%$ and a $\omega^+ = 10000$; and DQ500971, belonging to the subcluster 3, had a $q^+ = 1.6\%$ and a $\omega^+ = 292.969$. Finally, for *Nitrosotalea* cluster, BranchSiteREL reported 4 branches under EDS. AB457631 with a $q^+ = 9.2\%$ and a $\omega^+ = 66.511$; EF207220 with $q^+ = 3.4\%$ and $\omega^+ = 48.836$; Node703 with $q^+ = 3.7$ and $\omega^+ = 41.045$; and DQ148793 with $q^+ = 2.5\%$ and $\omega^+ = 18.792$.

Differential selection test

We compared every *amoA* cluster (Figure 5.4A), habitat (Figure 5.4B), cluster specific habitat (Figures 5.5A and 5.5B) and cluster specific sediment habitats (Figures 5.6A and 5.6B) to analyze the existence of different selection pressures on individual sites for internal and terminal branches with a $p \leq 0.05$.

Differential selection test results (internal/terminal) for main *amoA* clusters comparisons revealed 37/52 codons as selected differentially for the pair *Nitrosopumilus* and *Nitrososphaera* (Tables SC1 and SC21); purifying selection strength was slightly stronger in *Nitrosophaera*. Comparisons for *Nitrosopumilus* and *Nitrosotalea* (Tables SC2 and SC22) resulted in 15/18 codons and 19/24 codons for *Nitrososphaera* and *Nitrosotalea* (Tables SC3 and SC23). The trends observed in the values of selection strength were similar for both pairs of comparisons combined codons with stronger purifying selection {31, 98, 199} and others with a strong diversifying strength {46, 131}.

When all habitats were compared, differential selection test for marine habitat identified 7/11 codons with marine sediment habitats (Tables SC9 and SC29); 7/10 codons with estuarine sediments (Tables SC11 and SC31); 9/19 codons with hot springs (Tables SC13 and SC33) and 16/19 codons with soils (Tables SC6 and SC26). Marine sediments had 3/2 codons with estuarine sediments (Tables SC8 and SC28); 6/20 codons with hot spring (Tables SC10 and SC30) and 16/22 codons with soils (Tables SC4 and SC24). Estuarine sediment had 6/14 codons with hot spring (Tables SC12 and SC32); 13/14 codons with soils (Tables SC5 and SC25). Hot springs had 12/10 codons with soils (Tables SC7 and SC27).

Within cluster habitat specific alignments, tests for differential selection reported for *Nitrosopumilus* in marine habitats 5/16 codons with marine sediment (Tables SC16 and SC38) and 10/10 codons with estuarine sediment (Tables SC17 and SC39); marine sediment had 1/4 codons with estuarine sediment (Tables SC20 and SC35). *Nitrososphaera* in soil had 2/0 codons in marine sediment (Tables SC18) and 2/1 codons in estuarine sediment (Tables SC19 and SC40); marine sediment and estuarine sediment had 2/0 codons selected differentially (Table 34).

In addition, we did a codon-based maximum likelihood reconstruction

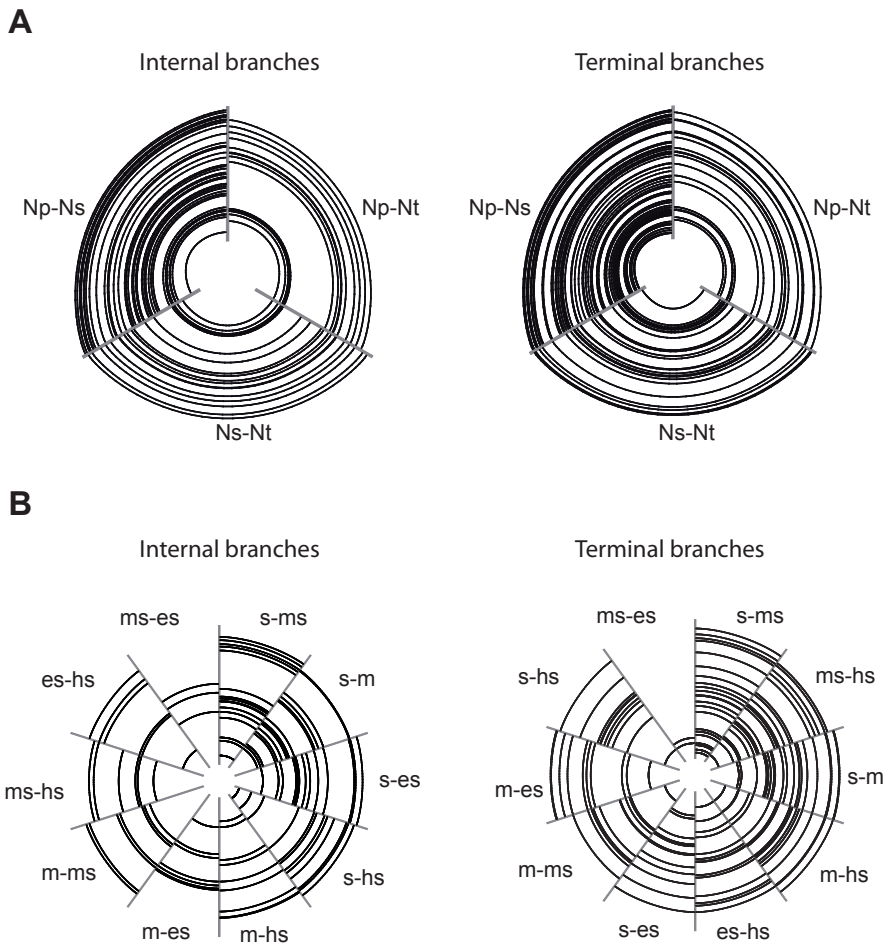

Figure 5.4: Hive plot with the distribution of differentially evolving sites (internal and external branches) detected at $p \leq 0.05$, on *amoA* clusters (A) and habitats (B). Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 form the genomic sequence of *Nitrosopumilus*; codon 1 located at the axis inner part of the hive plot. Ns:*Nitrosopumilus*; Np=*Nitrososphaera*; Nt=*Nitrosotalea*; m=marine; ms=marine sediment; es=estuarine sediment; hs=hot spring; s=soil.

of evolutionary history at the codon positions differentially selected in internal branches for the pair marine sediment and estuarine sediment in *Nitrosopumilus* and in *Nitrososphaera* (Figures S8-10 and Tables SD4-6). In *Nitrosopumilus*, we observed a different behavior for the patterns of substitution in codon {82} along the tree when we compared the two habitats; marine sediment, showed non-synonymous substitutions along internal branches as well as on terminal, while in estuarine sediment, mostly non-synonymous

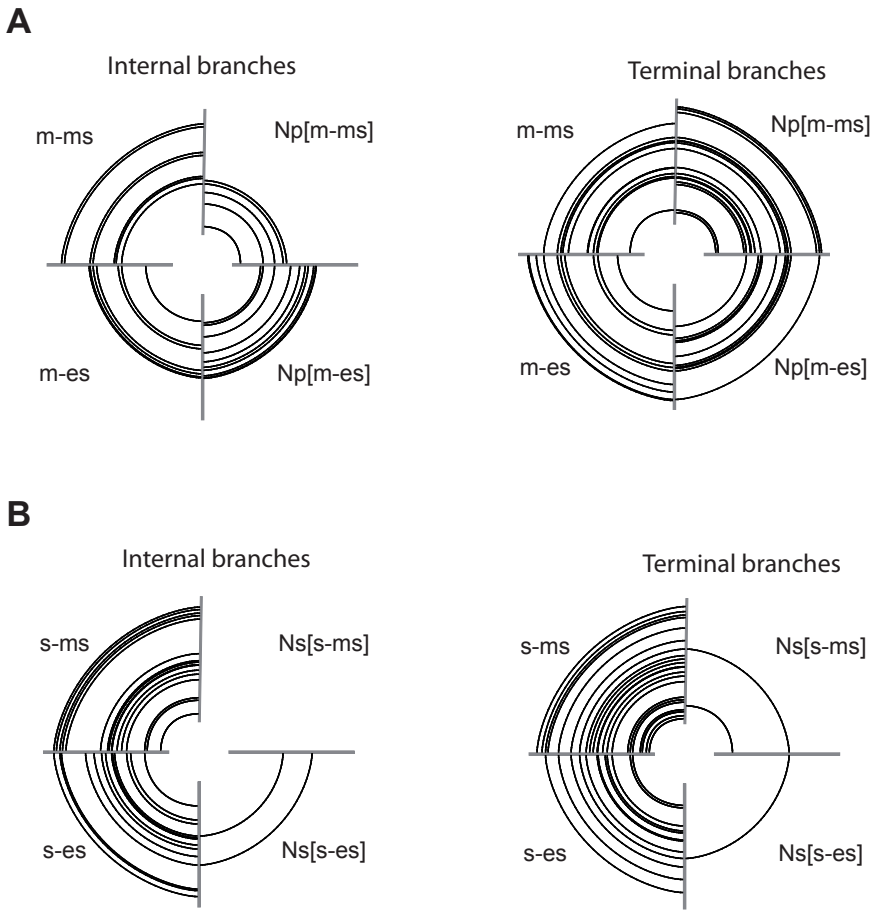

Figure 5.5: Hive plot with the distribution of differentially evolving sites (internal and external branches) detected at $p \leq 0.05$, on *amoA* cluster specific habitats for *Nitrospumilus* (A) and *Nitrososphaera*. Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 form the genomic sequence of *Nitrosopumilus*; codon 1 located at the axis inner part of the hive plot. Ns: *Nitrosopumilus*; Np=*Nitrososphaera*; Nt=*Nitrosotalea*; m=marine; ms=marine sediment; es=estuarine sediment; s=soil.

and synonymous substitutions were found on terminal branches only. In *Nitrososphaera*, codon {47} had similar pattern in both habitats, synonymous substitutions along internal and terminal branches; while codon {187} in estuarine sediment, non-synonymous and synonymous substitutions were found on terminal branches only but not in marine sediment.

Furthermore, we compared the common habitats in clusters, marine sediment and estuarine sediment, *Nitrosopumilus* had 10/27 codons (Tables SC15

and SC37 and *Nitrososphaera* had 21/22 codons differentially selected (Tables SC14 and SC36).

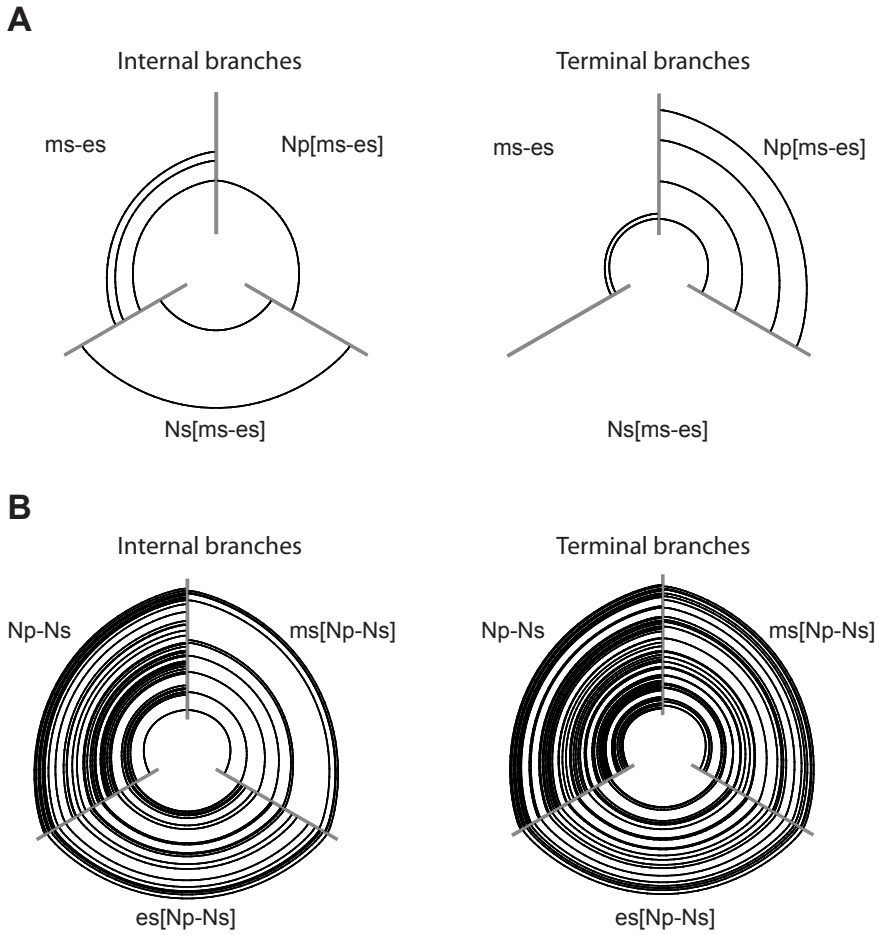

Figure 5.6: Hive plot with the distribution of differentially evolving sites (internal and external branches) detected at $p \leq 0.05$, on *amoA* cluster specific sediment habitats for *Nitrospumilus* (A) and *Nitrososphaera*. Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 form the genomic sequence of *Nitrosopumilus*; codon 1 located at the axis inner part of the hive plot. Ns: *Nitrosopumilus*; Np=Nitrososphaera; Nt=Nitrosotalea; m=marine; ms=marine sediment; es=estuarine sediment; hs=hot spring; s=soil.

To recognize which differentially selected codons in common habitats were caused by cluster differences or by habitat influence, we crosschecked differentially selected codon positions from the pair *Nitrosopumilus* and *Nitrososphaera* with the ones from the cluster specific common habitats. In marine sediment, 23 of 27 codons, and in estuarine sediment, 17 of 22 codons,

had also been differentially selected in cluster alignments terminal branches. Four codons {20, 48, 147, 178} in marine sediment and five codons in estuarine sediment {66, 70, 101, 116, 178} could be inferred as differentially selected by habitat influence. When we checked on internal branches, we got 8 of 10 codons in marine sediment; and 14 of 21 codons in estuarine sediments. Codons {122, 195} in marine sediment and codons {66, 101, 116, 122, 146, 166, 186} in estuarine sediment could differentially selected by habitat adaptation effects.

In *Nitrososphaera*, codons {47,187} are evolving differentially between habitats. Codon {47} is under a stronger purifying selection in both habitats while codon {187} it is in estuarine sediments. In *Nitrosopumilus*, populations from marine sediment and estuarine sediment differs only in codon {82}; selection strength, as *normalized* $dN - dS$, in estuarine sediments is -1.49 while in marine sediment is -0.35. Ancestral state reconstruction at this codon agrees with the fact this site is highly conserved on estuarine sediments by purifying selection. To complete the picture, we added marine habitats (planktonic) to *Nitrosopumilus* comparisons. Marine with marine sediments have 5 different evolving codons {14, 53, 73, 88, 94}; while with estuarine sediments has 10 codons {49, 53, 73, 101, 116, 126, 131, 141, 144, 145} and signal of purifying selection is stronger in marine samples.

When we do the test between clusters, we detect the intrinsic different evolving codons between clusters plus the habitat specific. When we isolate those codons we are able to identify codons {122, 195} in marine sediment and codons {66, 101, 116, 122, 146, 166, 186}. Codon {122} and codons {66, 101, 116, 122, 166, 186} are under stronger purifying selection on *Nitrososphaera* while codon {195} and codon {146} in *Nitrosopumilus*

5.4 Discussion

In the present work we analyzed the molecular evolution patterns of archaeal ammonia oxidizers linked to their ecological information unveiling which evolutionary processes and individual codons were behind the processes of diversification and adaptation. We are able to identify individual codons that underlie the factors responsible for habitat partitioning on the major *amoA* clusters from the consensus phylogeny, as they were subjected to differential selection and evolve differentially between habitats. The three archaeal ammonia oxidizing clusters analysed have different life styles and are found in very different environments. While *Nitrosopumilus* cluster members are generally found in open ocean and adapted to live in oligotrophic environments, *Nitrososphaera* and *Nitrosotalea* clusters are ammonia oxidizing archaeons from soil and acidic soils and usually nitrogen rich environments. To address the

question about what factors are driving *amoA* diversification we started our analyses at the cluster level (global) to capture the selective pressures at the taxonomic level leaving in the background the environmental signal (mixture of habitats). Then, we approached the other way around, splitting the phylogeny in five general habitats (Table 5.1), we wanted to maximize the effect of habitats in front of the taxonomic effect. And finally, we selected cluster specific habitats; therefore we are able to minimize the taxonomic effect associated to the habitat.

Values of mean dN/dS in all alignments indicate that the *amoA* gene is under purifying selection, deleterious mutations are removed from the AOA population, probably due functional constraints. However, we found strong evidences of episodic diversifying selection for individual sites and for lineages as well. MEME consistently identified episodes of diversifying evolution affecting a small subset of branches at individual sites, where site methods often report purifying selection at the same site. In the case of *amoA* we observed that episodic selection is widespread and we can conclude that the number of sites experiencing positive selection may have been underestimated Biller et al. (2012).

While *Nitrosopumilus* arised as the cluster with more codons under diversifying selection, and, clear differences were observed with *Nitrososphaera*, in terms of numbers and distribution of sites (Figure 5.1A). *Nitrososphaera* sites under EDS are grouped in four separated regions of the *amoA* gene, whereas *Nitrosopumilus* had a more parsimonious distribution along the gene. Not only in individual sites *Nitrosopumilus* has more events of EDS, also in the number of lineages; Figure 4 shows how lineages under EDS are spread around the *Nitrosopumilus* subtree, having a representative in almost every subcluster.

Marine sediment and estuarine sediment habitats had almost exactly the same number of sites under EDS, half of them found sharing the same position.. Different evolutionary histories combined with specific diversifying codons are the source behind the diversification and adaptation processes observed, e.g. codon 57 in estuarine sediments, presented non-synonymous and synonymous substitutions located on terminal branches only, while in marine sediment are substitutions were found throughout the tree. Similar conclusions can be extracted when we analyze cluster specific habitat.

When we analyzed the selective pressures at the population level (internal branches) we found many codons with non-synonymous substitutions only at the tips. Those sites had lower variation than the ones with both internal and terminal substitutions, suggesting that many recent substitutions were removed by long-term purifying selection. The same process was also observed in soil habitat, many codons were found only in tips, while the other habitats included in this study had more internal and terminal non-synonymous sub-

stitutions. Those results associated to habitats suggests that *amoA* sequences from all habitats but soil can be under a more active diversification process that can result in a higher capability for adaptation with a consequent increase of the species-level diversity. Recently, it has been suggested that natural selection could be basically episodic with transient periods of adaptive evolution that are masked by the prevalence of purifying or neutral selection on other branches (Murrell et al., 2012). This diversification mediated by bursts of positive selection followed by extensive conservation (homogenizing selection) shows evidences of the generation and maintenance of an evolutionary AOA *seed bank* driving the adaptive radiation into the different *amoA* phylogenetic clusters and ultimately in so different habitats (Martiny et al., 2006). This behaviour is in agreement with the AOA diversification patterns observed in the lineage through time plots in Fernàndez-Guerra & Casamayor (2012) where diversification events are followed by a stasis interval until a new burst in diversification.

As we have shown, *amoA* has an enormous potential for diversification and selective pressures are acting in different ways on the *amoA* phylogenetic clusters during the adaptive process. We were able to detect which codons were under different selective pressures and were evolving differentially between clusters, unveiling the key components for adaptation, as those codons were under evolutionary pressures for change (diversifying selection) or conservation (purifying selection) to maintain their biological fitness. The differences we detect at cluster level are the result of the evolutionary history of the ancestral *amoA* diversification and how evolutionary forces drives their cladogenesis determining the actual phylogenetic relationships. At the population level -internal branches test-, differences in sites evolving differentially between *Nitrosopumulis* and *Nitrososphaera* are striking; more than 25% of the codons are under a different selection regimes. Interestingly 86% of the sites evolving differentially between *Nitrosopumilus* and *Nitrosotalea* are also in *Nitrososphaera* and *Nitrosotalea*. The lower number of different evolving codons between *Nitrosotalea* and *Nitrosopumilus* agrees with their closer phylogenetic relationship.

A similar approach can be used to explore the habitat partitioning observed within *amoA* clusters, we analysed the differential selection between habitats. Two recent studies have shown that the environmental factors that influences *amoA* sequence diversity in marine environments are salinity, temperature and depth (Biller et al., 2012) and in terrestrial habitats soil pH (Gubry-Rangin et al., 2011). We have focused on the differences related with salinity in marine habitats comparing the estuarine sediments against the marine sediments. *Nitrosopumilus* and *Nitrososphaera* showed a different behaviour in terms of selection when we compared the cluster specific sed-

iment habitats. From our results we can infer that the estuarine and marine sediments in *Nitrososphaera* come from allochthonous inputs based on the low number of differential evolving sites. However, in *Nitrosopumilus* we observed a real adaptation to these habitats probably based on the saline gradient existing between the sediments from estuaries and open oceans. In this case our results suggests that *amoA* protein-coding sequence have a proportion of sites affected by different selective regimes in each habitat.

In conclusion, using a molecular evolution approach we were able to shed some light on the diversification processes of AOA, we established a hypothesis for the generation and maintenance of the *amoA seed bank* based on the evidences of observed evolutionary events; we identified the putative codons involved on *amoA* cladogenesis; and we were able to identify putative codons involved in the adaptation processes that could have a role in AOA habitat partitioning within clusters. Furthermore, with the sites we detect being under EDSor those which are evolving differentially, we provided a set of targets to develop or improve primers design for specific populations. Finally, this pionnering work will show its full potential once the crystallized structure of the archael *amoA* is available, helping for a better understanding of the physiological implications derived from the differences observed at the molecular level.

Acknowledgements

We thank Ramiro Logares for the computing time in the Barcelona Supercomputing Center and to the Centre de Supercomputació de Catalunya for their supercomputing facilities; we also thank Sergei L. Kosakovsky Pond for his assistance in the HYPHY analyses.

Appendix [2].

[2]See more Supplementary Information http://nodens.ceab.csic.es/ecoevo/ch5/

6

Looking for AOA Distribution by Fingerprinting Analysis in Marine Environments

Resumen

Los Polimorfismos de longitud de fragmentos de restricción (T-RFLP) es una técnica utilizada para analizar comunidades microbianas complejas. Permite la cuantificación de los filotipos más abundantes y se ha utilizado principalmente para comparar diferentes comunidades. T-RFPred ha sido desarrollado para identificar y asignar información taxonómica a los picos de los cromatogramos obtenidos en los T-RFLP, para poder realizar una descripción más intensiva de las comunidad microbianas. El programa estima el tamaño esperado de los 16S rRNA representativos para un determinado cebador y enzima de restricción y proporciona una asignación taxonómica.

6.1 T-RFPred nucleotide sequence size prediction tool for microbial community description based on terminal-restriction fragment length polymorphism chromatograms

Abstract [1]

Background

Terminal-Restriction Fragment Length Polymorphism (T-RFLP) is a technique used to analyze complex microbial communities. It allows for the quantification of unique or numerically dominant phylotypes in amplicon pools and it has been used primarily for comparisons between different communities. T-RFPred, Terminal-Restriction Fragment Prediction, was developed to identify and assign taxonomic information to chromatogram peaks of a T-RFLP fingerprint for a more comprehensive description of microbial communities. The program estimates the expected fragment size of representative 16S rRNA gene sequences (either from a complementary clone library or from public databases) for a given primer and restriction enzyme(s) and provides candidate taxonomic assignments.

Results

To show the accuracy of the program, T-RFLP profiles of a marine bacterial community were described using artificial bacterioplankton clone libraries of sequences obtained from public databases. For all valid chromatogram peaks, a phylogenetic group could be assigned.

Conclusions

T-RFPred offers enhanced functionality of T-RFLP profile analysis over current available programs. In particular, it circumvents the need for full-length 16S rRNA gene sequences during taxonomic assignments of T-RF peaks. Thus, large 16S rRNA gene datasets from environmental studies, including metagenomes, or public databases can be used as the reference set. Furthermore, T-RFPred is useful in experimental design for the selection of primers as well as the type and number of restriction enzymes that will yield informative chromatograms from natural microbial communities.

[1] See original publication in Fernàndez-Guerra et al. (2010).

6.1.1 Background

Terminal-Restriction Fragment Length Polymorphism (T-RFLP) analysis of 16S rRNA gene amplicons is a rapid fingerprinting method for characterization of microbial communities (Liu et al., 1997; Marsh, 1999). It is based on the restriction endonuclease digestion profile of fluorescently end-labeled PCR products. The digested products are separated by capillary gel electrophoresis, detected and registered on an automated sequence analyzer. Each T-RF is represented by a peak in the output chromatogram and corresponds to members of the community that share a given terminal fragment size. Peak area is proportional to the abundance of the T-RF in the PCR amplicon pool, which can be used as a proxy for relative abundance in natural populations (Blackwood et al., 2003).

This method is rapid, relatively inexpensive and provides distinct profiles that reflect the taxonomic composition of sampled communities. Although it has extensively been used for comparative purposes, a T-RFLP fingerprint alone does not allow for conclusive taxonomic identification of individual phylotypes because it is technically challenging to recover terminal fragments for direct sequencing. However, when coupled with sequence data for representative 16S rRNA genes, T-RF identification is feasible (e.g. Gonzalez et al. (2000); Mou et al. (2005); Pinhassi et al. (2005)). Here we describe a method to assign the T-RF peaks generated by T-RFLP analysis with either 16S rRNA gene sequences obtained from clone libraries of the same samples, metagenome sequences or data from public 16S rRNA sequence databases. T-RFPred can thus be used to classify T-RFs from T-RFLP profiles for which reference clone libraries are not available, albeit with lower phylogenetic resolution, by taking advantage of the wealth of 16S rRNA gene sequence data available from metagenome studies and public databases such as the Ribosomal Database Project (RDP) (Cole et al., 2007) or SILVA (Pruesse et al., 2007).

Metagenome sequencing studies from a variety of environments are accumulating at a rapid pace. While most often partial gene sequences, these libraries have the advantage that they are less subject to biases of other PCR-based techniques (see e. g. (Kanagawa, 2003) for a review) and, thus, can better represent the original community structure. Furthermore, both metagenome and pyrosequencing of tagged 16S rRNA gene amplicons provides unprecedented coverage of 16S rRNA gene diversity in specific environments. Therefore, these types of datasets are valuable references when attempting to taxonomically classify T-RF peaks from diverse microbial communities.

Tools have been previously developed to perform in silico digestions of 16S rRNA gene sequences and/or to assign a taxonomic label to the chromatograms. Such programs include TAP-TRFLP (Marsh et al., 2000), MiCA (Shyu et al., 2007), T-RFLP Phylogenetic Assignment Tool (PAT) (Kent

et al., 2003), TReFID (Rosch & Bothe, 2005), TRAMPR (Fitzjohn & Dickie, 2007), an ARB-software integrated tool (Ricke et al., 2005) and TRiFLe (Junier et al., 2008). Table 6.1 contains some of the essential features of these packages. The most obvious advantage of T-RFPred as compared with other available software applications is that the program handles either partial or full-length user input sequences. This is because T-RFPred retrieves complete sequences of close relatives from the public databases for T-RF assignments and at the same time it taxonomically bins the clone sequences. Furthermore, it can use large sequence datasets of virtually any size as reference sets in taxonomic assignments. T-RFPred is exclusive to 16S rRNA gene sequences and designed to exploit the full potential of T-RFLP profiles and their use in the description of prokaryotic communities.

6.1.2 Implementation

T-RFPred is coded in Perl and uses the BioPerl Toolkit (Stajich et al., 2002), fuzznuc from the EMBOSS package (Rice et al., 2000) and the BLASTN program from the NCBI BLAST suite (Altschul et al., 1990). T-RFPred has been tested in Unix-like environments, but runs in all the operating systems able to execute Perl, BioPerl, BLAST and EMBOSS; a ready-to-use VMware virtual image is also available for download at http://nodens.ceab.csic.es/t-rfpred/.

An interactive shell guides the user through the multiple steps of the analysis. Users can choose to analyze archaeal or bacterial sequences using either forward or reverse primers. The primer search utilizes fuzznuc, which allows the user to select the number of nucleotide ambiguities. The program extracts a subset of sequences from the RDP database that will supplement sequence analysis of clone libraries. T-RFPred generates and exports in a tab delimited text file: (1) the fragment length for the RDP sequence with the best BLASTN hit to the input sequence(s), (2) the estimated fragment length for the input sequence, (3) the gap length for the input sequence, (4) the percent identity between the input sequence and the best hit RDP sequence and (5) the taxonomic classification. The BLASTN search results and the Smith-Waterman alignments (Smith & Waterman, 1981) are saved to allow the user to manually check the results.

Database

The program uses a custom version of the aligned RDP as a flat file in FASTA format, where the header has been modified to include the NCBI taxonomic information and the forward/reverse position of the first non-gap character from the RDP alignment. T-RFPred exploits the Bio::DB::Flat capabilities from BioPerl to index the RDP flat file for the rapid retrieval of 16S rRNA gene

Table 6.1: Characteristics of the available software to assign a phylogenetic label to the chromatogram fragment peaks

Software package	Characteristics	Reference
TAP-TRFLP	Web-based. Although it can be accessed through the older version of the Ribosomal Database Project, it has not been updated.	Marsh et al. (2000)
MiCA	Web-based. Newest version (MiCA 3) allows the selection of primers and in silico digestion of database sequences. Does not allow for user input sequences.	
T-RFLP Phylogenetic Assignment Tool (PAT)	Web-based. Contains database of terminal restriction fragment sizes. Allows for the upload of fragment size database.	Kent et al. (2003)
TReFID	Downloadable. Databases include 16S rRNA gene, dinitrogenase reductase gene (nifH) and nitrous oxide reductase gene (nosZ). Limited number of sequences although the user could expand it.	Rosch & Bothe (2005)
TRAMPR	R package. Based on a database of known T-RFLP profiles that can be constructed by the user. Loads data directly from ABI output files. Allows analysis with any type of gene, primer set and restriction enzyme.	Fitzjohn & Dickie (2007)
ARB-software integrated tool (TRF-CUT)	Part of the ARB software. Allows for user input sequences that need to be aligned before analysis. Any type of gene could be analyzed.	Ricke et al. (2005)
TRiFLe	Java based. Allows for user input sequences. Can analyze any type of gene.	Junier et al. (2008)
T-RFPred	Handles large database, such as 16S rRNA sequences from metagenomes, of user input clone sequences that do not need to be full length; multiple platforms. Makes use of the Ribosomal Database Project sequence database, which updates regularly. User needs to install Perl, Bioperl, BLAST and EMBOSS.	This study

Complete sequence at least at the 5′-end of the sample sequence is needed in every case except for T-RFPred, as this program finds the closest related sequence in the Ribosomal Database Project database by BLASTN.

sequences. All restriction enzymes available in REBase (Roberts et al., 2005) are stored in a flat file and available for use in the analysis. A list of frequently used forward and reverse primers is available, although the user may also input custom primers.

Algorithm

In part, the rationale for the described method was to circumvent the need for full-length 16S rRNA gene sequences from representative clone libraries. In addition to requiring multiple sequencing reactions, obtaining full-length sequences is generally complicated by the ambiguous nature of the 5′ end of a sequence generated by the Sanger approach (i.e. the first 10-30 bp of a sequence are missing). When the same primer set used to generate T-RFLP profiles is also used to generate amplicons for libraries and directional sequencing of representative clones, as is often the case, in silico predictions of expected peak sizes are cumbersome. Additionally, the size of the fragment is subject to experimental error (Kaplan & Kitts, 2003; Marsh, 2005), which complicates the assignment of chromatogram peaks to specific phylogenetic groups. T-RFPred takes advantage of the most comprehensive database of 16S rRNA gene sequences (the RDP) to identify the closest related sequences for analysis to provide more definitive phylogenetic assignments of chromatogram peaks. Collectively, the Perl scripts achieve the following steps:

1. Create a subset of all the sequences in the RDP with nucleotide information spanning the region targeted by the fluorescently labeled primer and with a length > 1200 nucleotides for Bacteria and > 900 nucleotides for Archaea.

2. Convert the subset created in Step 1 into a BLAST-ready database using formatdb. Conduct a BLASTN search with the sample sequences (FASTA format) against the RDP database and extract the best hits.

3. Determine if sample sequences have the denoted restriction enzyme recognition site. If the cut site is present, proceed to Step 4. If the cut site is not present, estimate the expected fragment size using the closest RDP sequence and proceed to Step 5.

4. Generate a Smith-Waterman alignment of the sample sequence with the best hit from the RDP. This will provide accurate percent identities and the start/end positions of the alignment needed to estimate the fragment sizes.

5. Obtain the position of the restriction enzyme recognition site in the aligned sample sequence and the primer position in the RDP sequence.

Use the RDP sequence to calculate the number of nucleotides in the gap between the primer and the start position of the Smith-Waterman alignment as shown in Figure 6.1.

6. Assign a taxonomic classification using the best RDP BLAST hit.

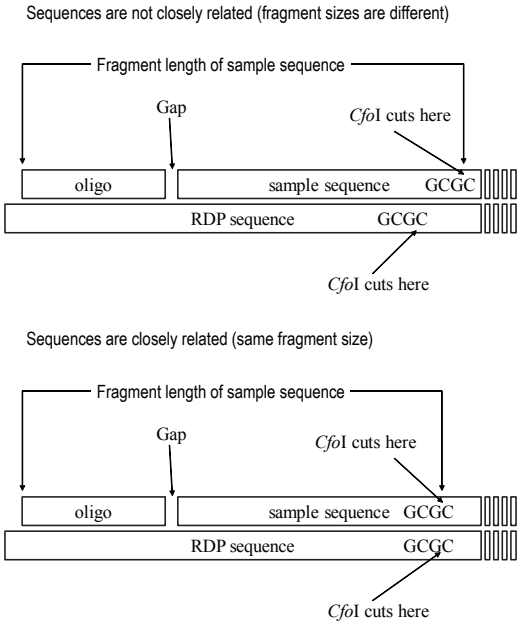

Figure 6.1: Description of the method to estimate the length of the terminal-fragment size for partial 16S rRNA sequences.
The closest sequences (by homology search) in the RDP database are used to estimate the length of the fragment and its phylogenetic affiliation. The primer sequence is fluorescently labeled and it is close to the 5' end of the 16S rDNA gene. 'Gap' is the missing part of the sequence between the position of the primer and the beginning of the sequence. The position of the target sequence determines the size of the terminal fragment.

6.1.3 Results and Discussion

We have developed a computational method to provide putative phylogenetic affinities of chromatogram peaks of 16S rRNA gene T-RFLP profiles. Additional file 1, Supplementary Tables S1-S3 show the typical output of T-RFPred for the clone sequences from Gonzalez et al. (2000), Mou et al. (2005), and Pinhassi et al. (2005), respectively. The T-RFPred output provides the estimated fragment size of the digested clone sequences as well as a user defined number

of closest relatives. This feature is valuable for estimating the conservation of the digested product size for a given enzyme and taxonomic group analyzed.

T-RFPred was also evaluated by reanalyzing chromatogram peaks from T-RFLP profiles of marine communities described in (Gonzalez et al., 2000). Two 16S rRNA datasets constructed from sequences from public databases, designated "4926" (4926 bacterioplankton Genbank sequences) and "GOS" (6370 Global Ocean Sampling Expedition Microbial Metagenome sequences; (Rusch et al., 2007)), were analyzed with T-RFPred using three restriction enzymes (i.e., CfoI, HaeIII, and AluI). Details on experimental procedure are described in the Additional File 1. The two datasets and their predicted fragment sizes and phylogenetic affiliations were used to taxonomically label the chromatogram peaks from natural samples (Figure 6.2).

With very few exceptions, all valid fragment peaks were properly identified and in good agreement with the phylogenetic assignments reported in the literature using complementary clone libraries (Table 2). For instance, from the 4926 sequence dataset analyzed with three restriction enzymes, 124 clones yielded in silico digested fragment sizes matching peaks labeled as "1" (previously identified as *Alphaproteobacteria* of the *Roseobacter* clade) in Figure 6.2. Of these clones, 90% (111 clones) were properly classified as Roseobacter-related, seven were *Alphaproteobacteria* outside the *Roseobacter* group, four *Gammaproteobacteria*, and two were *Betaproteobacteria* (Table 6.2). Thus, these T-RFs were labeled as *Roseobacter*. Those peaks labeled with a "2" (Figure 6.2) were mapped to members of the SAR11 group as 119 of the 148 sequences (80%) were from this lineage (Table 6.2). The chromatogram peak assignments were less ambiguous when the GOS dataset was used as the reference. With regards to T-RFs labeled 1 and 2 in Figure 6.2, 95% of the sequences belonged to the *Roseobacter* group and all (n = 269) sequences belonged to the SAR11 group (Table 6.2). Therefore, the GOS dataset was more representative of the diversity of the bacterioplankton in the natural samples. This might be because that dataset was comprised of sequences exclusively from surface seawater samples; the T-RFLP profiles analyzed were also generated from surface seawater.

6.1.4 Conclusions

T-RFLP is a popular method for analysis of microbial communities and in silico automated methods are needed to facilitate the taxonomic identification of T-RFs in community profiles. Traditionally, computational methods to analyze T-RFLP experiments follow one of two approaches: (a) in silico simulation of the digestion of reference sequences from databases to find the most suitable enzymes that describes the microbial community organization or (b)

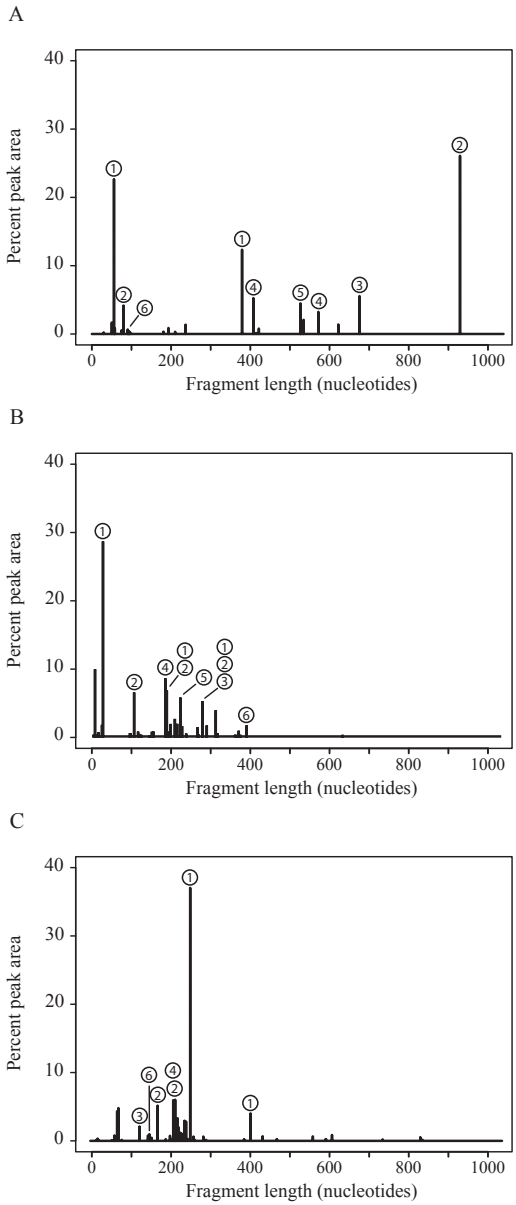

Figure 6.2: Evaluation of the T-RFPred prediction tool.
Graphics of terminal fragment profiles generated from (A) CfoI, (B) HaeIII, and (C) AluI restriction enzymes digestions of 16S rDNAs amplified from total community DNA as described in Gonzalez et al. (2000). The taxonomic affiliations for the numerical labels are as follows: 1, *Roseobacter*; 2, *SAR11*; 3, *Cyanobacteria*; 4, *SAR86*; 5, *SAR116*; and 6, *SAR324*.

Table 6.2: Phylogenetic information for the 16S rRNA sequences present in the 4926 and GOS datasets that matched selected chromatogram peaks shown in Figure 6.2

Dataset	Peak	Chromatograms	Number of sequences	Taxonomic group
4926	1	CfoI, HaeIII	243	Total
			146	*Roseobacter*
			53	*Gammaproteobacteria*
			36	*Alphaproteobacteria*
			4	*Planctomycetes*
			2	*Betaproteobacteria*
			1	*Cyanobacteria*
			1	*Firmicutes*
4926	1	CfoI, HaeIII, AluI	124	Total
			111	*Roseobacter*
			7	*Alphaproteobacteria*
			4	*Gammaproteobacteria*
			2	*Betaproteobacteria*
4926	2	CfoI, HaeIII	207	Total
			152	*SAR11*
			51	*Firmicutes*
			1	*Alphaproteobacteria*
			1	*Unclassified Bacteria*
4926	2	CfoI, HaeIII, AluI	148	Total
			119	*SAR11*
			29	*Firmicutes*
GOS	1	CfoI, HaeIII	263	Total
			231	*Roseobacter*
			18	*Alphaproteobacteria*
			13	*Gammaproteobacteria*
			1	*Actinobacteria*
GOS	1	CfoI, HaeIII, AluI	243	Total
			229	*Roseobacter*
			12	*Alphaproteobacteria*
			1	*Gammaproteobacteria*
			1	*Actinobacteria*
GOS	2	CfoI, HaeIII	560	Total
			559	*SAR11*
			1	*Alphaproteobacteria*
GOS	2	CfoI, HaeIII, AluI	269	Total
			269	*SAR11*

Sequences that matched the fragment sizes were analyzed using 2-3 different restriction enzymes as indicated. *Alphaproteobaceria* sensu latu refers to any bacterial sequences in the class that were not either *Roseobacter* or *SAR11*. See Experimental Procedures in the Additional File 1 for details.

T-RF from experiments can be binned to the in silico generated fragments to identify the taxonomic groups present in the sample. T-RFPred is designed to provide a list of candidate taxa that corresponds to the chromatogram peaks using a complementary reference clone library or public databases.

Depending upon the restriction enzyme used, broad phylogenetic groups can sometimes give the same fragment size. Thus, we also determined that community profiles generated with at least two different restriction enzymes are needed for the most robust taxonomic identifications (Table 2). The method has also its caveats as is not meant to positively identify phylogenetic groups or species based upon terminal fragment length, particularly, as the identification of the sequences cannot be solely determined based on the closest BLASTN hit alone. Manual inspection of the BLASTN hits and additional efforts may also be needed for more conclusive taxonomic assignments. In the example above, we conducted homology searches (BLASTN) to a set of reference sequences from representative taxa as well as phylogenetic treeing methods to confirm the taxonomic affiliations of the GOS and 4926 sequences whose predicted fragment sizes matched a chromatogram peaks (data not shown). Despite these caveats, the position of restriction enzyme recognition sites within the 16S rDNA molecule does reflect a level of phylogeny and can be used to help guide experimental design (i.e. which and how many restriction enzymes are most appropriate for a given community) so that the most reliable results for the T-RFLP characterization of a given prokaryotic assemblage can be obtained.

In summary, T-RFPred offers an alternative, freeware and open source program for researchers using T-RFLP to examine microbial populations. The program can help researchers determine the most appropriate restriction enzyme(s) to use when designing experiments to assess community structure using the T-RFLP method. It can also provide information on the taxonomic assignments of specific T-RFs without the need for comprehensive complementary clone libraries.

Acknowledgements

This work was supported by grant PIRENA CGL2009-13318-CO2-01/BOS to EOC, grant CTM2007-63753-C02-01/MAR to JMG, and grant CONSOLIDER-INGENIO2010 GRACCIE CSD2007-00067 to AFG from the Spanish Ministry of Science and Innovation, and grant OCE-0550485 from the National Science Foundation to AB.

Availability and requirements

- Project name: T-RFPred

- Project home page: http://nodens.ceab.csic.es/t-rfpred/
- Operating systems: Linux (tested in Debian, Ubuntu and RHEL), Mac OS X (tested in MacOS X 10.5 and Mac OS X 10.6), Windows (via a Xubuntu VMware image)
- Programming language: Perl
- Other requirements: BioPerl, BLAST and EMBOSS
- License: none
- Any restrictions to use by non-academics: none

Appendix [2]

[2] See more Supplementary Information in Fernàndez-Guerra et al. (2010).

Annex

Targeted metagenomics by community fingerprinting

Next Generation Sequencing (NGS) is an expanding field and every year there are new advances in the sequencing technologies developed, lowering the price of the Mb. Although this seems to be a democratization of the sequencing technologies, not every laboratory has the funds to sequence any environment searching for their gene or species of interest. So in this case, a targeted approach to choose the most suitable environment for sampling is essential. Terminal Restriction Fragment Length Polymorphism (T-RFLP) is a molecular biology technique for profiling microbial communities that has been shown to be really useful during the last years for the identification and quantification of microbial communities. Also T-RFLP could be complemented with clone libraries as an alternate way to obtain sequence information about particular microorganisms in a community profile. A first screening using T-RFLPs from the potential sampling site could increase the chances of success to find the functions we are interested in.

As an example to explain combination of community fingerprinting and metagenomics, we selected metagenomes from the Global Ocean Sampling Expedition (GOS) dataset (Rusch et al., 2007) to test the performance of 16S rRNA gene community fingerprinting on crenarcheal ammonia oxidizers. Two different samples were examined, the freshwater lake Gatun in the Panama Channel and in Coastal marine sites. For that purpose, we performed a virtual digestion of the 16S rRNA genes from these sampling sites and a posterior taxonomic assignment of the predicted fragments using T-RFPred software (Fernàndez-Guerra et al., 2010). On Table 6.3 there are the size of the peaks for each digested fragment; in Figure 6.3A we have the taxonomic affiliations of the 16S rRNA gene performed by (Yilmaz et al., 2012) and in Figure 6.3B there are the community profiles for each sampling

Sample	Group	RsaI	HhaI	MspI	AfaI
Coastal	Eury	11, 252, 262, 607	-	19, 147	79, 251, 261, 606
	Thau	-	261	-	-
Fresh water	Eury	-	-	-	-
	Thau	26, 41, 602	326, 1297	283	26, 40, 601

Table 6.3: Peak size from the fragment resulting from the virtual digestion shown in Figure 6.3. Eury: *Euryarcheota*; Thau: *Thaumarchaeota*

sites after performing the digestion using three different restriction enzymes. The fingerprinting showed presence of thaumarchaeota in both samples. In the marine site we observed euryarchaeal peaks prevailing, whereas thaumarchaeota prevailed in the freshwater site, in agreement with the picture obtained from the relative abundances of the sequencing dataset. A preliminary analysis using a fingerprinting method like T-RFLP of our suspected sites would improve the information we can obtain from our metagenomic studies.

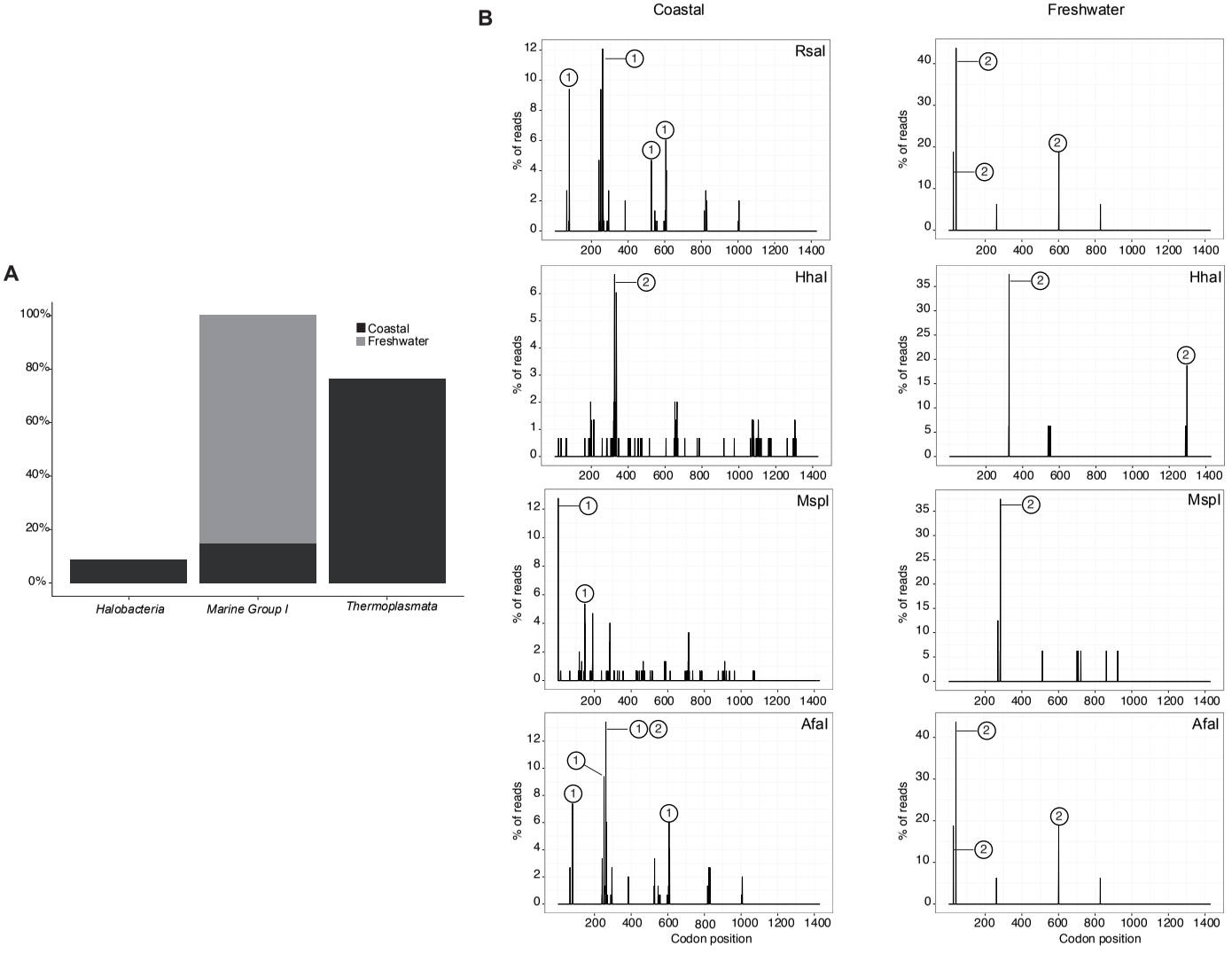

Figure 6.3: A) Taxonomic affiliations of the 16S rRNA gene by (Yilmaz et al., 2012) and B) Graphics of terminal fragment profiles generated from RsaI, HhaI, MspI and AluI restriction enzymes by virtual digestion of 16S rRNAs genes using T-RFPred software. The taxonomic affiliations for the numerical labels reported by T-RFPred are as follows: 1, Marine Group II *Euryarchaeota*; 2, Marine *Thaumarchaeota*

7

An evolutionary perspective on the phylogenetic partitioning in marine ammonia-oxidizing *Thaumarchaeota*

Resumen

Las *Thaumarchaeota* marinos son unos contribuyentes importantes en las primeras etapas de la nitrificación en los oceanos y además tienen un importante papel en les ciclos biogequímicos marinos. Estos *Thaumarchaeota* marinos presentan dos grupos diferentes correspondientes a un ecotipo de superficie (*shallow*) y otro de aguas profundas (*deep*). Esta partición observada se cree que es originada por adapataciones relacionadas con procesos de fotoinhibición-resistencia y adaptación a la presión hidrostática. En el siguiente trabajo analizamos los diferentes ecotipos para determinar las presiones selectivas que han actuado sobre el gen de la *amoA*. Hemos encotrado al codon {89} como un potencial componente clave en las adaptaciones a la luz. Por otro lado, hemos encontrado evidencias que el gen de la *amoA* se encuentra bajo intensa selección a nivel molecular para adaptarse a las condiciones de las profundidades marinas. Por último, hemos encontrado que el linaje donde diversifican los dos ecotipos están sujetos a selección episódica diversificadora; posiblemente como resultado de la rápida diversificación y radiación adaptativa que este grupo ha experimentado en la columna de agua marina.

Abstract [1]

Marine *Thaumarchaeota* are important contributors to the first step in oceanic nitrification (ammonia oxidation) and have an important role on marine biogeochemical cycles. These Marine *Thaumarchaeota* presents a two phylogenetically distinct clusters corresponding to surface (*shallow*) and deep water (*deep*) ecotypes. The partition observed has been suggested to be the result of photoinhibition-resistance adaptations and hydrostatic pressure adaptations. In the present work we analyze the different Marine *Thaumarchaeota* ecotypes to find where the selective pressures acted in the adaptation of *amoA* gene. We found that codon {89} as a potential key player in the light adaptation process. We were able to identify the branching point between the *shallow* and *deep* group to be under EDS. We found evidences that the *amoA* gene is under intense selection to be adapted at the molecular level to the environmental conditions prevailing in the deep sea. And we found signal of episodic diversifying selection on the branching point between the *shallow* and *deep* ecotypes, possibly as result from the rapid diversification and adaptive radiation that this group experienced in the marine water column.

7.1 Introduction

Marine *Thaumarchaeota* (formely Marine Group I *Crenarchaeota*) are important contributors to the first step in oceanic nitrification (ammonia oxidation) and have an important role on marine biogeochemical cycles. In several studies (Francis et al., 2005; Hallam et al., 2006; Mincer et al., 2007; Beman et al., 2008; Santoro et al., 2010; Hu et al., 2011a; Mosier & Francis, 2011; Biller et al., 2012) have been reported a two phylogenetically distinct clusters corresponding to surface (*shallow*) and deep water (*deep*) ecotypes. This significant phylogenetic partitioning has been related to photoinhibition-resistance adaptations (Mincer et al., 2007), as the amo genes are known to be membrane spanning and this enzymatic complex could experience significant exposure to light. For many years, this photoinhibitory effect was known in the ammonia oxidizing bacteria (AOB) (Hooper & Terry, 1973; Horrigan & Springer, 1990; MA & RD, 1996a,b) and recently has been studied extensively by (Merbt et al., 2012) in archaeal ammonia oxidizers (AOA). There are evidences of species-specific and dose- wavelength-dependent photoinhibition and contribute to niche differentiation between and within AOA and AOB determining their distribution and diversity in light-affected ecosystems.

Hydrostatic pressure adaptations also have been suggested as a factor

[1] Fernàndez-Guerra, A, EO Casamayor. **Manuscript in preparation.**

that could influence the ecological and phylogenetic partitioning. Measurable piezophilic adaptations generally occur at deeps greater than 500-1000 m (Yayanos, 1995).Some Another factor interesting to analyze in the column water is the oxygen deprivation found in the OMZ. Some marine AOA have preference for environments of low (<10 μM) levels of dissolved O_2, where the ammonia oxidation might be coupled to anaerobic ammonium oxidation (anammox) and/or denitrification (Molina et al., 2007; Lam et al., 2007; Beman et al., 2008; Lam et al., 2009; Bouskill et al., 2011; Pitcher et al., 2011; Yan et al., 2012; Ulloa et al., 2012).

To unveil the processes and mechanisms underlying the different adaptation processes observed we analyzed the different vertical marine ecozones in order to (i) find signatures of diversifying and episodic selection at the codon level; (ii) detect the lineages subject to episodic diversifying selection; and (iii) compare selection patterns along the vertical gradient in the marine realm.

7.2 Methods

Sequence collection and environmental annotation

The planktonic archaeal *amoA* sequences used in the present work were retrieved from NCBI GenBank database release 191 (August 2012) using the search string "amo subunit A or ammonia monooxygenase subunit A or ammonia monooxygenase α subunit or Amo α subunit or amoA OR ammonia monooxygenase A or ammonia monooxygenase or ammonium monooxygenase or ammonium monooxygenase A) NOT (genome or chromosome or plasmid or bacteria[ORG])" as previously described in (Fernàndez-Guerra & Casamayor, 2012). We retrieved a total of 22,284 sequences. Then we proceeded to perform a series of filtering steps that comprises a validation of the annotation of the gene using HMMER 3.0 (Eddy, 2010) in combination with the PFAM 26.0 (Punta et al., 2012) model PF12942 for the archaeal *amoA* domain and the removal of the sequences that lacked the *isolation_source* tag and contained any ambiguities. At the end, we finished with a data set of curated 3,019 sequences. Then we clustered at the 97% identity (Pester et al., 2012) with usearch6 (Edgar, 2010b). Finally, we ended with 366 representatives OTUs used for the posterior analyses.

Using the information from the *isolation_source* tag and the publication associated when needed, we classified sequences depending on depth according to the following three marine vertical zones: 174 OTUs from the surface euphotic layer (100 m depth); 126 OTUs from the oxygen minimum zone (250 and 750 m depth), and 66 OTUs from bathy- and abyssopelagic depths (1750–7000 m depth). The euphotic layer OTUs corresponded to the *shallow*

ecotype, and the sequences from the OMZ and the bathy- abyssopelagic zone to the *deep* ecotype from Marine *Thaumarchaeota*. We used the *amoA* gene directly obtained from the genome *Nitrosopumilus maritimus SCM1* (YP_001582834.1) as the reference sequence to refer codon positions and the genomic *amoA* sequence from Candidatus *Nitrososphaera gargensis* Ga9.2 (YP_006863114.1) as an outgroup for the phylogenetic inference.

Sequence alignment

To achieve a high alignment quality for further codon-based analyses, first, we checked whether selected sequences were on frame doing a BLAST search against their respective amino acid sequences using *bl2seq* with the algorithm tblastn from NCBI Blast+ 2.2.25 package (Camacho et al., 2009). Then, the nucleotide sequences were spliced following the coordinates of the blast results. Later, to minimize the codon alignments errors and improve the detection of the selection events (Jordan & Goldman, 2011; Privman et al., 2011) we used an in-house modified version of GUIDANCE (Penn et al., 2010). This customized version of GUIDANCE takes profit of parallelization in some of the steps required by the GUIDANCE algorithm and uses RAxML 7.3.0 (Stamatakis, 2006) as fast maximum-likelihood phylogenetic method to infer the guiding trees for MUSCLE (Edgar, 2010a). The high quality codon alignment after applying the filtering algorithms implemented in GUIDANCE resulted in 585 positions with a confidence score to each individual position.

Screening for recombination and phylogenetic inference

Recombination events may mislead selection analysis as each partition could have different rates. To correct for this effect, selection codon based analysis has to take into account a tree topology for each partition (Scheffler et al., 2006). We did a GARD screening on 5 different random data sets picking up 20 sequences from each vertical marine ecozone. We repeated GARD analysis at least five times for each random data set to check convergence and test whether the recombination breakpoints found were stable due to the aleatory nature of GARD.

To use the implementation of the Kishino-Hasegawa test (KH) (Kishino & Hasegawa, 1989) bundled in HYPHY to test the congruence of topologies, we inferred a ML tree for each partition and for each ecozone using RAxML 7.3.0. We peformed1000 maximum likelihood searches with 1000 bootstrap replicates to find the best-scoring tree under the GTRGAMMA model for each partition of the alignment determined by the breakpoint location and using

the confidence scores obtained by GUIDANCE. For each topology we calculated the sitewise log likelihoods. To check that the incongruences found by the KH test were caused by a different ratio in each side of the breakpoint, *p-values* of each alignment combination according to the Approximately Unbiased (AU) test using multiscale bootstrapping (Shimodaira, 2002) and the Shimodaira-Hasegawa (SH) test (Shimodaira & Hasegawa, 1999) were calculated using CONSEL (Shimodaira & Hasegawa, 2001). Tree topologies for every marine vertical ecozone inferred for each partition were used for later codon based analyses. The distance matrices were calculated using the TN93 genetic distance.

Codon based analyses

We used HyPhy (Pond et al., 2005), a package for testing hypothesis using phylogenies, to detect signatures of positive and negative selection and differential adaptive evolution from the *amoA* codon alignments estimating the ratio of nonsynonymous (dN or β) and synonymous substitution rates (dS or α). First, we performed pairwise comparisons of genetic distance between vertical zones to be sure we were analyzing different populations and not polymorphisms (Kryazhimskiy & Plotkin, 2008). We calculated the distance based F_{ST} pairwise metrics using the distance methods implemented in HYPHY (Zárate et al., 2007) with 1000 bootstraps and 1000 permutations.

Then, we searched the best nucleotide model to be combined with the codon model MG94 (Muse-Gaut 1994). Nucleotide models that fitted best each data set by the AIC_C criterion could be found in Table 7.1.

To estimate the rates of nonsynonymous and synonymous substitutions and identify which codon sites were under pervasive negative selection we used the Fixed Effects Likelihood for terminal (FEL) and internal (iFEL) branches taking in account recombination (Pond & Frost, 2005; Pond et al., 2006).

To identify positive selection at the level of individual sites we applied the Mixed Effects Model of Evolution (MEME) (Murrell et al., 2012)a mixed effects model of evolution that allows the distribution of ω to vary from site to site and also from branch to branch at a site. Because MEME only detects individual sites, we applied the BranchSiteREL method (Pond et al., 2011) to detect which lineages were under episodic diversifying selection (EDS).

HyPhy provided the possibility to compare differential evolution in the different marine ecozones following the hypothesis that at a given codon site, the dN/dS ratio differed between two samples, on the entire tree and on the internal branches as described in (Pond et al., 2006). This method only detected sites subject to different selective pressures in both samples, regardless

Table 7.1: Sequences analyzed in this study, classified by layer. N=number of sequences; S=Tree length as expected expected substitution per codon site; Model: Best nucleotide model by AIC_c; MEME, FEL, iFEL, dN internal, dN Tips Only: number of codons detected at $p \leq 0.05$

Layer	N	S	Model	Mean dN/dS	MEME
Euphotic	174	6.93/6.88	012340	0.044	25
OMZ	126	3.95/3.43	012340	0.042	16
Bathy- Abyssopelagic	66	2.61/1.86	012310	0.051	6

Layer	FEL	iFEL	dN Intenal	dN Tips Only
Euphotic	185	183	112	5
OMZ	183	185	88	10
Bathy- Abyssopelagic	124	103	60	14

of which residue appeared to be selected for.

All codon-based analyses were run on the non-recombinant fragments to take account of recombination separately. To plot the lineages detected by BranchSiteREL to be under episodic diversifying selection we performed a phylogenetic inference for the whole data set with RAxML 7.3.0 with 1000 maximum likelihood searches and 1000 bootstrap replicates to find the best-scoring tree under the GTRGAMMA model for each partition of the alignment determined by the breakpoint location and using the confidence scores obtained by GUIDANCE. Then, we made a consensus tree form both topologies using the extended majority rule implemented in RAxML. Trees were graphically represented with iTOL (Letunic & Bork, 2007).
In order to deal with convergence problems, most analyses were run at least twice with different initial conditions, and the run with the best likelihood was kept.

7.3 Results

Recombination signals

The runs on the five random data sets of GARD algorithm reported breakpoints at positions 323, 335, 344, 353 and 359. Thus, recombination signal detected by GARD suggested a breakpoint nearby position 343. We therefore partitioned the *amoA* gene alignments at the breakpoint position and we inferred a maximum likelihood tree for each partition as described in the Methods section Then, we tested the inferred tree topologies for incongruences using the KH test implemented in HYPHY, in all cases KH test reported a

$p < 0.01$ supporting the presence of a breakpoint. In addition, CONSEL reported *p-values < 0.01* for AU and SH tests.

Codon based analyses

Pairwise comparisons of genetic distance between ecozones, calculated by the F_{ST} statistic, showed that the euphotic layer diverged significantly from the deeper layers at $p < 0.001$. While, the results for the OMZ and the bathy-abyssopelagic ecozone were not significant ($p = 0.27$). And the tree lengths (S) calculated as the expected substitution per codon site along the tree (Table 5.1) assured medium sequence divergence (Anisimova et al., 2002). Accurate prediction in similar sequences is possible when we include a large number of lineages. For example a $S = 6.93$ in euphotic zone partition means an average branch length of $S/(2T-3) \approx 0.02$ nucleotide substitutions per codon (where T are the number of taxa).

Mean ω values (Table 7.1) suggested that the *amoA* gene sequences from the different marine ecozones were under the effects of purifying selection. When searching for individual sites under purifying selection, FEL (terminal branches) detected 181 sites in the euphotic zone, 165 in the OMZ and 146 in the bathy- abyssopelagic zone, whereas IFEL (internal branches) detected 179 sites in the euphotic layer, 161 in the OMZ and 128 in the bathy- abyssopelagic layer. Overall, neither FEL nor iFEL detected any site under positive selection.

One of the advantages of IFEL is the estimation of dN and dS separately in both internal and the terminal branches. Thus, it is possible to identify if recent nonsynonymous substitutions (terminal branches) were not represented on internal branches or the other way around (Table 7.1). We observed differences in the ratio of nonsynonymous substitution on terminal branches and on internal branches (terminal:internal) was 5:112 in the euphotic zone, 10:88 on the OMZ and 14:60 on the bathy- abyssopelagic zone. Therefore, in all three ecozones there were more codons with nonsynonymous substitutions in internal branches (population level) than in the terminal branches, suggesting local adaptation of the AOA, where transient substitutions are removed by purifying selection.

We identified signatures of episodic or diversifying selection (EDS), through MEME the strength and the proportion of branches that were under episodic diversifying selection (q^+) with a $p \leq 0.05$ (Figure 7.1; Table S7-9). For the AOA inhabiting the euphotic zone, 12 codon sites were detected to be under EDS, 1 in the OMZ, and 8 in the bathy- abyssopelagic layer (Figure 1). Several codon sites under EDS where shared among the AOA inhabiting the different ecozones (Figure 7.1). For instance, codon {188} in the euphotic zone ($\beta^+/\alpha = 39.58$; $q^+ = 1.6\%$) and in the OMZ ($\beta^+/\alpha = 506.72$;

$q^+ = 2.8\%$);codon {58} with $\beta^+/\alpha = 66.17$ and $q^+ = 6.4\%$ in the euphotic zone and $\beta^+/\alpha = 11.4$ and $q^+ = 6.0\%$ in the bathy- abyssopelagic zone; codon {123} with $\beta^+/\alpha = 39.58$ and $q^+ = 6.4\%$ in the euphotic zone and $\beta^+/\alpha = 139.54$ and $q^+ = 1\%$ in the bathy- abyssopelagic zone; and codon {192} with $\beta^+/\alpha = 27.07$ and $q^+ = 1.8\%$ in the euphotic zone and $\beta^+/\alpha = 71.52$ and $q^+ = 3.4\%$ in the bathy- abyssopelagic zone. Interestingly, AOA in the euphotic zone showed 8 specific EDS codon sites {43, 44, 54, 100, 107, 108, 188, 199, 201} and 5 the bathy- abyssopelagic zone {25, 53, 90, 147, 194}.

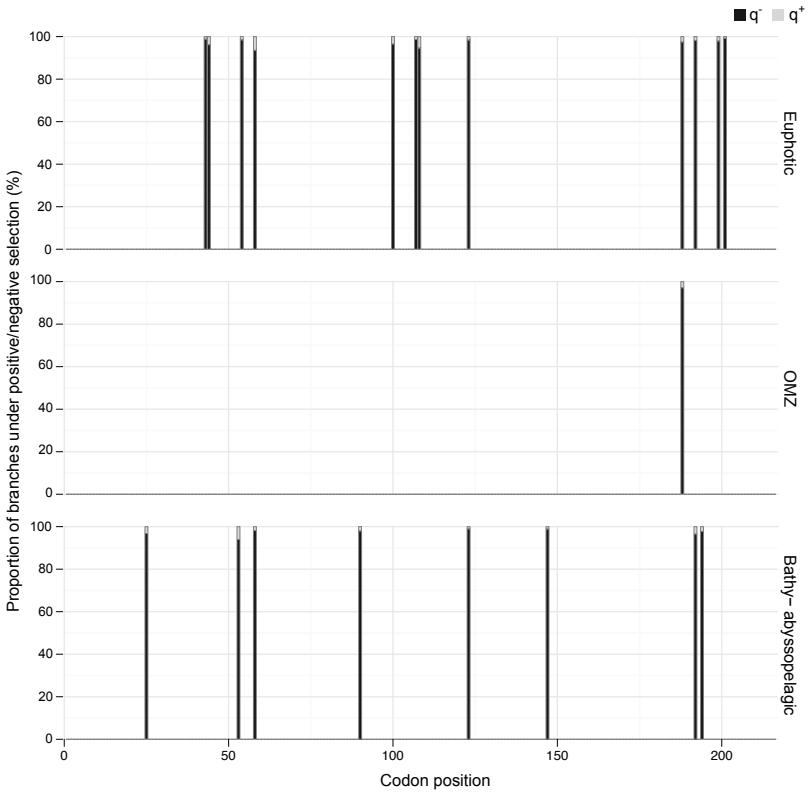

Figure 7.1: Distribution of sites detected, $p \leq 0.05$, under Episodic Diversifying Selection by MEME for every marine layer. Codon positions based of the *amoA* aminoacid sequence YP_001582834.1 from the genomic sequence of *Nitrosopumilus maritimus SCM1*. q^-: proportion of branches under purifying selection; q^+: proportion of branches under diversifying selection.

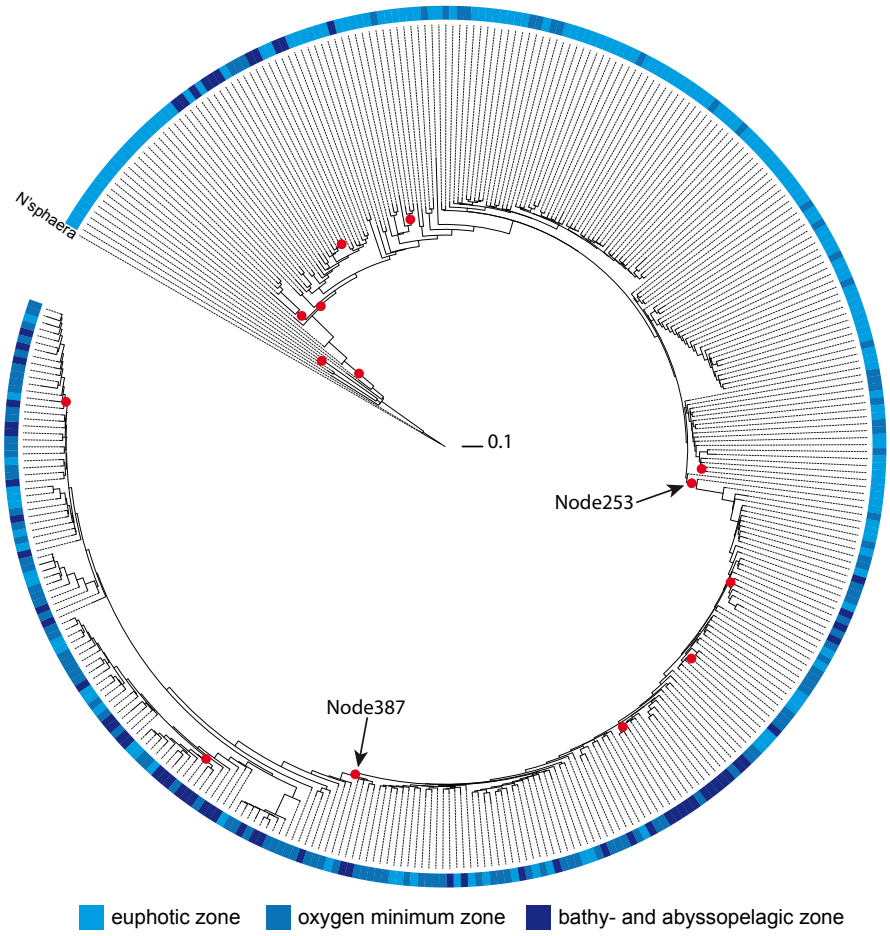

Figure 7.2: Maximum likelihood inferred phylogeny for the Marine Group I *Crenarcheota*. Branches detected under episodic diversifying selection by the sequential test at $p \leq 0.05$ are marked with a red circle. Each tree is scaled on the expected substitution per codon site. Candidatus *Nitrososphaera gargensis* Ga9.2 (YP_006863114.1) is the outgroup.

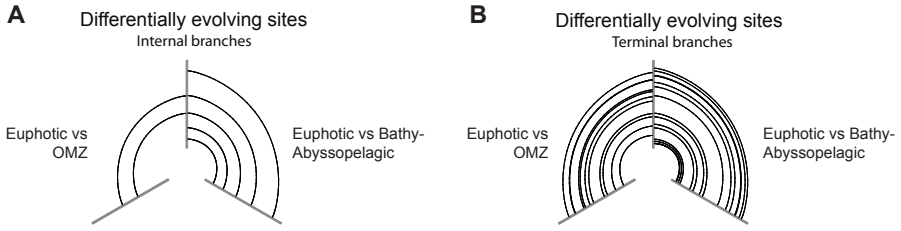

Figure 7.3: Hive plot with the distribution of differentially evolving sites on the different marine layers at $p \leq 0.05$. Codon positions based of the amoA aminoacid sequence YP_001582834.1 form the genomic sequence of *Nitrosopumilus maritimus SCM1*; codon 1 located at the axis inner part of the hive plot.

Lineages under episodic diversifying selection

BranchSiteREL method found evidences of lineages under episodic diversifying selection with a corrected $p \leq 0.05$ (Holm's procedure). Detailed results for the branch-level mixture of negative, (nearly) neutral and positive selection models values are shown in Table S16. The analysis reported evidences of 14 lineages where the strength of diversifying selection (ω^+)> 500 and the proportion of sites under selection (q^+) varied from 0.7 to 10.1%. The tree topology (Figure 7.3) revealed a clear partitioning in two clusters. One dominated by AOA sequences from the euphotic layer (*shallow* ecotype, split in 7 lineages) and the other constituted mainly by AOA sequences from the OMZ and the bathy- abyssopelagic zone (*deep* ecotype, split in the remaining 7 lineages). Two of the lineages detected are particularly interesting in the diversification process of the planktonic AOA, e.g., Node253 ($\omega^+ = 3333.11$; $q^+ = 3.3\%$) and Node387 ($\omega^+ = 1559.51$; $q^+ = 1.1\%$). Node253 corresponded to the branching event between the *shallow* ecotype and the *deep* ecotype in Marine *Thaumarchaeota*.

Comparing selection pressures in different marine ecozones

We compared the sequences from the AOA *shallow* ecotype against the *deep* ecotype to analyze the existence of different selection pressures. Looking at individual sites at the population level (internal branches), AOA from the euphotic zone showed two differentially evolving codon sites {89, 132} with the OMZ and five codon sites {25, 53, 89, 132, 194} with the bathy- abyssopelagic layer at a $p \leq 0.05$.

However, when terminal branches were compared, the euphotic zone showed eleven differentially evolving codon sites {34, 57, 80, 89, 119, 130, 143,

147, 171, 172, 190} with the OMZ and fifteen codon sites {14, 15, 20, 25, 53, 62, 80, 89, 132, 155, 166, 182, 183, 194, 201} with the bathy- abyssopelagic zone at $p \leq 0.05$. Thus, we observed that codon {89} evolved differentially in *shallow* and *deep* ecotypes. Codon 89 may potentially be related to the adaptation to the photoinhibitory processes observed in the euphotic zone (Church et al., 2010; Merbt et al., 2012). We performed a maximum likelihood ancestral state reconstruction for the codon {89}, as this codon was evolving differentially in the internal branches and in the terminal branches. In the euphotic layer the values of the *normalized* $dN - dS$ for internal and terminal branches were smaller than those in the deeper zones (TableS10, TableS11, TableS13 and TableS14) indicating the effect of purifying selection to maintain the residue on the euphotic zones. In Figure S1, Figure S2 and Figure S3 there are the different evolutionary history of codon {89} for the euphotic zone, the OMZ and the bathy- abyssopelagic zone respectively. In agreement with the *normalized* $dN - dS$ values observed, in the euphotic zone there were mainly synonymous substitutions along the internal and terminal branches, while in the deeper layers nonsynonymous substitutions dominated its evolutionary history (Table S17).

7.4 Concluding remarks

In the present work we carried out a molecular evolutionary analysis to explore the mechanisms involved in the depth-dependent phylogenetic partitioning observed in the marine ammonia oxidizers *Thaumarchaeota*. This partitioning has been observed in terms of microbial community structure and phylogenetic analyses in multiple studies (Mincer et al., 2007; Yakimov et al., 2007; Varela et al., 2008; De Corte et al., 2008; Agogué et al., 2008; Yakimov et al., 2009; Kalanetra et al., 2009; De Corte et al., 2009; Church et al., 2010; Bouskill et al., 2011; Hu et al., 2011b; Sintes et al., 2012). To explain the factors driving this partitioning, (Mincer et al., 2007) suggested photoinhibition-resistance and piezophilic adaptations as the main forces. As a result of these adaptations, particular residue substitutions could be affected. To find the signatures of selection in the partitioning and adaptation to each marine ecozone, we compared available sequences of the amoA gene from thaumarchaeotal populations from the *shallow* ecotype (euphotic) against the populations from the *deep* group (OMZ and bathy- abyssopelagic). We split the populations from the *deep* group in the populations from the OMZ and the populations from the bathy- abyssopelagic layer, because measurable piezophilic adaptations generally occur at depths greater than 1000 m (Yayanos, 1995). OMZ are usually detected above such threshold.

Ammonia oxidation is a pivotal energy-gaining pathway required to sus-

tain metabolic activity in the different vertical zones we analyzed (de la Torre et al., 2008; Könneke et al., 2005; Yakimov et al., 2009).). As a result of the importance of this pathway, *amoA* gene experiences functional constraints reflected on the signatures of purifying selection detected. Although amoA gene is under purifying selection, it is known that natural selection is basically episodic and is mediated by bursts of selection localized in a subset of sites and in a small number of lineages (Murrell et al., 2012) allowing the diversification and adaptation of the different *amoA* phylotypes. ending in the diversification and adaptation of the different amoA phylotypes. In all three marine vertical ecozones we have been able to detect signatures of episodic diversifying selection. In the euphotic zone, the codon sites under EDS are concentrated in three regions along the gene. As the set of AMO genes are known to be membrane spanning and the enzymatic complex could experience significant exposure to light, the regions that contains the codon sites detected to be under EDS could be the result to the adaptation for the high light conditions from the surface waters producing changes in the protein structure that confers a better resistance to the photoinhibition observed in AOA (Beman et al., 2008; Merbt et al., 2012). Thus hypothesis certainly deserves further investigations.

Regarding the *deep* group, the F_{ST} statistic indicated that the OMZ and the bathy- abyssopelagic zone hadn't diverged enough and are the same microbial population. In this scenario, one expects similar behaviors in terms of the EDS signatures between the two ecozones. However, the analyses revealed a high level of adaptation and functional constrain in the populations inhabiting the OMZ. The almost inexistent sites under EDS in the OMZ are related to sustained stability and the specific environmental properties of the OMZ. In the OMZ there is a sharp decrease of dissolved O_2 ($< 10\mu$M) and a characteristic vertical distribution of nitrogen compounds, e.g. NO_3^- deficit, a secondary nitrite maxima, and low NH_4^+ concentrations along the vertical gradient, as well as N2O maxima towards the OMZ boundaries (Molina et al., 2007; Beman et al., 2008; Lam et al., 2009; Bouskill et al., 2011; Pitcher et al., 2011; Yan et al., 2012; Ulloa et al., 2012). Such strong chemical gradients could be considered gene flow barriers and their oxygen-deficient waters as suitable environments for innovation, where organism develop specific adaptations to exploit these unique conditions (Rogers, 2000). Unlike the OMZ *amoA* genes, the bathy- abyssopelagic zone *amoA* genes have more codon sites under EDS, indicating that adaptation processes are undergoing on the deeper layers of the water column, most probably due the piezophylic and temperature adaptations (Biller et al., 2012).

In the analysis to explore the differential adaptation in the different marine layers we were able to identify a codon site {89} as a potential key player

in the light adaptation process. Photoinhibition appears to affect the amoA, and Merbt et al. (2012) demonstrated that the effect of photoinhibition is significantly greater in AOA than AOB. In aquatic environments, AOA presence increases with depth, when light intensity decreases (Church et al., 2010), but, although this evidences of photoinhibition, AOA *amoA* abundance is high in regions with high irradiance, such as surface waters (Galand et al., 2010) and high mountain lakes (Auguet et al., 2011; Auguet & Casamayor, 2008). Church et al. (2010) found a high transcriptional activity as a response to the photoinhibition in the upper water column in the Pacific Ocean, where the abundance of AOA was low; but the high abundance of AOA in the previously described high irradiance environments, could support our findings about specific adaptations at the molecular level. The combination of the conservation of codon {89} with codons {43, 44, 54, 100, 107, 108, 188, 199, 201} detected to be under EDS, can be another factor to be included in the explanation of the adaptive process of AOA to the high irradiance conditions of marine surface.

When we explored the adaptation of the *amoA* gene to the high pressures conditions found in the bathy- abyssopelagic layer compared with the euphotic layer, most of the codon sites detected to be differentially selected in the bathy- abyssopelagic were under EDS. This finding could be interpreted as an evidence that the *amoA* gene is under intense selection to be adapted at the molecular level to the environmental conditions prevailing in the deep sea.

Finally, in order to complete the evolutionary picture of the partitioning mechanisms of the marine *amoA* gene we analyzed the lineages to specifically find those under EDS. We were able to identify the branching point between the *shallow* and *deep* ecotypes to be under EDS. The signal of EDS on this lineage might have resulted from the rapid diversification and adaptive radiation that this group experienced in the marine water column. AOA are flexible enough to adapt to a wide range of environments but, unfortunately, it is difficult to assign a direction for the evolutionary processes observed because the origin of the archaeal ammonia oxidation is unknown. However, we have shown that the amoA is certainly a dynamic gene in terms of evolutionary processes, despite its functional constrains related to the energetic metabolism in Thaumarchaeota. These findings adds another view on the ecology of AOA centered on molecular evolutionary processes, that will help to understand the habitat partitioning and ecological success observed in worldwide distributed Thaumarchaeota.

Acknowledgements

We thank Ramiro Logares for the computing time in the Barcelona Supercomputing Center and to the Centre de Supercomputació de Catalunya for their supercomputing facilities; we also thank Sergei L. Kosakovsky Pond for his assistance in the HYPHY analyses.

Appendix [2]

[2]See more Supplementary Information in http://nodens.ceab.csic.es/ecoevo/ch7/

8

A Network Approach to Explore the Archaeal Ammonia Oxidation in Oceans trough Metagenomics

Resumen

El análisis del genoma de Candidatus *"Nitrosopumilus maritimus"* SCM1, un *Thaumarchaeota* marino, ha revelado la existéncia de un sistema diferente para la oxidación del amonio del descrito para bacterias. Actualmente se barajan dos hipótesis para explicar la falta del homólogo para la hidroxilamina oxidasa encontrado en bacterias. Unos sugieren que el nytroxyl puede ser utilizado como intermediario en lugar de la hidroxilamina; y otros que el proceso puede ser mediado por oxidasas presentes en el espacio periplásmico. En el siguiente trabajo hemos aplicado Modelos Gráficos Gaussianos para analizar la oxidación del amonio en arqueas. Hemos combinado el conocimiento de las familias de proteínas conocidas con la fracción proteica desconocida para encontrar asociaciones en terminos de función y estructura. Hemos sido capaces de determinar estas asociaciones a los genes de *Nitrosopumilus* y generar una serie de candidatos que podrían estar implicados en la oxidación del amonio.

Abstract [1]

The analysis of the marine AOA Candidatus *Nitrosopumilus maritimus* SCM1 genome revealed the existence of an ammonia oxidation system different from the one found in ammonia oxidizing bacteria (AOB); therefore two hypotheses had been proposed to explain the lack of the homolog for the hydroxylamine oxidase complex found in Bacteria. One hypothesis assumes that it could may driven by one of the periplasmic multicopper oxidases; and the other suggests the use of nytroxyl as intermediate instead of hydroxylamine. In the present work we applied Graphical Gaussian Models to analyze the archael ammonia oxidation in the Global Ocean Survey (GOS) metagenomic samples. We combined the knowledge of the known protein domain families with the *unknown unknowns* to discover *unknown* protein associations that lead us to understand better archaeal ammonia oxidation in terms of function and structure. We have been able to track the protein families to the specific genes in *Nitrosopumilus* genome and generate candidates, that later could be further explored by experimental assays, which by their special structural conformation, could be components for the ammonia oxidation itself or participate in the electron transport.

8.1 Introduction

Since metagenomic surveys in seawater (Venter et al., 2004) and soil (Treusch et al., 2005) revealed the existence of a different *amoA* gene related to the phylum *Thaumarchaeota*, the idea of a possible ammonia-oxidizing metabolism in widespread archaea was quickly expanded. Ammonia oxidizing arcahaea (AOA) had been widely detected in a large variety of aquatic and terrestrial environments (Rotthauwe et al., 1997; Nicol & Schleper, 2006; Francis et al., 2007; Agogué et al., 2008; Nicol et al., 2008; Auguet et al., 2010; Reigstad et al., 2008; Zhang et al., 2008; Auguet & Casamayor, 2008; Erguder et al., 2009; Tourna et al., 2011; Mussmann et al., 2011; Auguet et al., 2012; Pester et al., 2012). Ammonia oxidation is the first step of nitrification and is a biogeochemical process of global importance in natural worldwide ecosystems. Although the knowledge regarding the AOA biology has consistently increased during last years by the new AOA sequenced genomes (Walker et al., 2010; Kim et al., 2011; Tourna et al., 2011; Hallam et al., 2006) many aspects of the ammonia oxidation pathway still remain unclear. Recently the isolation of the marine AOA *Candidatus "Nitrosopumilus maritimus" strain SCM1* (Walker et al., 2010), un-

[1] Fernàndez-Guerra, A, A Barberán, R Kottmann, FO Glöckner, EO Casamayor. **Manuscript submitted.**

veiled the existence of an ammonia oxidation system based in a highly copper-dependent system for ammonia oxidation and electron transport, completely different from the present in ammonia oxidizing bacteria (AOB). AOA lack the homolog for the hydroxylamine oxidase complex found in AOB and, although it is not clear still how the process is carried out; (Walker et al., 2010) have proposed two hypothesesOne suggests that may occur via one of the periplasmic multicopper oxidases, and the other may involve nytroxyl as intermediate instead of hydroxylamine.

The ecology and environmental distribution of AOA has also been extensively studied in the most recent years, and a recent study has reported photoinhibition in the ammonia oxidizing activity of AOA, larger than in AOB. Several studies on AOA (Francis et al., 2005; Hallam et al., 2006; Mincer et al., 2007; Beman et al., 2008; Santoro et al., 2010; Hu et al., 2011a; Mosier & Francis, 2011; Biller et al., 2012) have reported the existence of two phylogenetically distinct AOA clusters corresponding to surface ("*shallow*") and deep water ("*deep*") ecotypes. To explain this significant phylogenetic partitioning (Mincer et al., 2007) proposed the existence of photoinhibition-resistance adaptations, as the amo genes are known to be membrane spanning and this enzymatic complex could experience significant exposure to light.

Large scale metagenomic surveys such as the Global Ocean Survey (GOS) (Rusch et al., 2007) have brought the chance to carry out a more holistic approach to study marine ecosystem (Karl, 2007) providing new clues to find missing pieces in biological pathways as the case of ammonia oxidation process in AOA. Metagenomics also increases the current limited view on the protein universe (Koonin, 2007; Gilbert et al., 2008) spectacularly expanding the number of both *known unknowns*, like the domains of unknown function (DUF), and *unknown unknowns*, putative coding sequences without any kind of previous knowledge (Harrington et al., 2007; Jaroszewski et al., 2009). Recently several approaches have been developed to try to assign functions to metagenomic fragments using a large repertoire of bioinformatic techniques including sequence homology searches, sequence clustering, self-organizing maps, support vector machines, relevance networks, and protein-protein interaction networks (Galperin & Koonin, 2004; Abe et al., 2009; Jaroszewski et al., 2009; Yooseph et al., 2008; Harrington et al., 2007; Hawkins et al., 2008). In the present work, we proposed a different approach to extract valuable information from the co-ocurrence of individual protein domains obtained from metagenomic complex datasets (Levin, 2003) using Graphical Gaussian Models (GGM). GGMs are very popular for modeling genomic data, i.e., reverse engineering of genetic regulatory networks, because they are able to distinguish direct from indirect interactions as it is explicitly considered the effect of all remaining observed variables (Schäfer & Strimmer, 2005a; Whit-

taker, 1990). Here, we combined the knowledge of the known protein domain families and 16S RNA gene with the *unknown unknowns* present in the GOS dataset to specifically looking for the archaeal ammonia oxidation process. The approach unveiled new associations of AOA with known protein families that helped for a better understanding of the associations of the known AMO with other biological processes. We also proposed candidate proteins to be further experimentally tested, which by their special structural conformation could be key components in the ammonia oxidation process in AOA or active participants in the electron transfer chains.

8.2 Methods

PFAM annotation

Unassembled reads from the GOS expedition (Rusch et al., 2007) were retrieved from the CAMERA database (Seshadri et al., 2007). We selected 53 surface water samples from picoplankton collected within the same size fraction (0.1–0.8 m), and free of bacterial contamination during sample handling (DeLong, 2005) The analyzed metagenomic data set comprised approximately 8000 Mb contained in approximately 5 million reads. We used HMMER 3 (Eddy, 2010) in combination with PFAM 26.0 (Punta et al., 2012) to screen the protein domains present on the six frame translation of the GOS reads with length ≥ 60 amino acids. We identified a hit significant if E-value ≤ 0.001 and bias is at the same order of magnitude as the sequence bit score. We followed such strict approach to avoid falsely significant non-homologous hits that merely shared a similar strong biased composition with the query model. In order to explore the archaeal ammonia oxidation process in our dataset, we reannotated *Nitrosopumilus maritimus* SCM1 genome (NC_010085) to update its functional annotation to PFAM release 26.0 as we used in the network inference.

Clustering of the *unknown* fraction

All the reads that were not assigned to a PFAM domain were clustered hierarchically using USEARCH 5.2 (Edgar, 2010b), first at 90% and then at 60% identity (Li et al., 2008a). We removed for posterior analyses all singleton clusters. We identified the reads encoding for 16S RNA and then we clustered those reads using USEARCH 5.2 at the 97% identity level((Gevers et al., 2005). We used the representative sequences from each cluster to retrieve their taxonomic annotation from SILVA database (Pruesse et al., 2007).

Network analysis

Data preparation: Filtering and normalization

We built abundance tables for the PFAM families, the clusters of *unknowns*, and for the annotated 16S rRNA gene clusters, respectively. We selected those PFAM families and 16S rRNA gene clusters that were present in at least 10% of the GOS sampling sites. In order to get more robust results we selected only those clusters of *unknowns* present in at least the 20% of GOS sampling sites. Next, read abundances counts were normalized. First we log-transformed the read counts; then we standardized the log-transformed values, for each transformed measure i of sample s, we subtracted the arithmetic mean of all transformed values and this difference was divided by the standard deviation of all log-transformed values for the given sample. Finally we performed a multiple sample scaling; all values were scaled from 0 to 1, and therefore we could remove the negative values derived from the log-transformation without affecting the relative differences within samples.

Graphical Gaussian models

For building a Graphical Gaussian Model network it has to be to estimated the empirical covariance matrix, calculate the partial correlation matrix and perform statistical tests to determine the partial correlation coefficients that were different from 0 and that correspond to the significant edges of the graph. After data normalization, abundance $tableX$ had $Nrows$ -number of sampling sites- and $Ccolumns$ -number of clusters, including PFAM families, *unknown clusters* and 16S rRNA gene clusters. X followed a multivariate normal distribution $\mathcal{N}_C = (\mu, \Sigma)$, with mean vector $\mu = (\mu_1, \ldots, \mu_C)$ and positive-definite covariance matrix $\Sigma = (\sigma_{ij})_{i \leq 1, j \leq C}$.

Covariance parameters σ_{ij} can be written as: $\sigma_{ij} = \rho_{ij} - \sigma_i \sigma_j$. Parameter ρ_{ij} corresponds to the Pearson correlation coefficient between clusters i and j; and σ_i^2 and σ_j^2 are the variance terms for clusters i and j. High correlation coefficients between two clusters in the Pearson correlation matrix $P = (\rho_{ij})_{(i \leq 1, j \leq C}$ could indicate direct associations, indirect associations or associations mediated by a common cluster. As we were interested only in direct associations, the partial correlation matrix $\tilde{P} = (\tilde{\rho}_{ij})_{1 \leq i,j \leq C}$ described the direct associations between two clusters whilst taking away the effects of another cluster, or several other clusters. Standard graphical modeling theory (Whittaker, 1990) showed that partial correlation matrix $\tilde{P}$ is related to the inverse of the covariance matrix σ (also known as the concentration matrix, Ω) leading to the straightforward estimator

$$\hat{\tilde{\rho}}_{i,j} = -\omega_{i,j} / \sqrt{\omega_{i,i} \omega_{j,j}},$$

where $\hat{\Omega} = (\hat{\omega}_{ij}) = \hat{\Sigma}^{-1}$ and $\hat{\omega}_{ij}$ is the $(i,j) - th$ element of $\hat{\Omega}$.

Our data settings had the effects of "small n, large p", the number p of clusters is much larger than the number n of sampling sites; this dimensionality issue affected the estimation of the unbiased empirical covariance matrix $\hat{S}$, with entries defined as

$$\hat{s}_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j),$$

S could not be considered a good approximation of the true covariance matrix as wasn't anymore positive-definite and couldn't be inverted as it became singular. Limitations caused by the "small n, large p" prevented a direct computation of the partial correlation. One way to overcome this limitation is to use shrinkage estimators of the covariance matrix. Schäfer & Strimmer (2005b) proposed a shrunk estimate of the covariance matrix using a James-Stein estimator. The idea was to construct a well-conditioned positive-definite matrix so that the matrix had full rank and could easily be inverted. The shrinkage estimator decreased $\hat{S}$ variance, and also reduced the mean squared error (MSE) by finding the best trade-off between error due to bias and error due to variance. The estimator shrinks the sample covariance matrix $\hat{S}$ towards a low-dimensional estimator $\hat{T}$ of the covariance matrix Σ. The linear shrinkage estimator Σ^* is defined as the linear combination of the estimators $\hat{S}$ and $\hat{T}$:

$$\Sigma^* = \lambda \hat{T} + (1 - \lambda)\hat{S},$$

where $\lambda \epsilon [0,1]$ represented the shrinkage intensity. There were several low-dimensional estimators of $\hat{T}$, for gene association networks, (Schäfer & Strimmer, 2005b) recommended to shrink the correlation terms towards zero and to leave the diagonal terms as estimated by the empirical variances. Shrinkage intensity could be calculated analytically:

$$\lambda = \frac{\Sigma_{i \neq j} \widehat{var}(s_{ij})}{\Sigma_{i \neq j} s_{ij}^2},$$

where s_{ij} are the empirical covariance parameters. Last step in the gene association network reconstruction was to assign statistical significance to the edges. For this purpose a mixture model,

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r}; \kappa) + (1 - \eta_0) f_A(\tilde{r}),$$

was fit to the observed partial correlation coefficients $\tilde{r}$ across edges. f_0 was the distribution under the null hypothesis of vanishing partial correlation, η_0 was the (unknown) proportion of "*null edges*", and f_A the distribution of observed partial correlations assigned to actually existing edges. The latter was

assumed to be an arbitrary nonparametric distribution that vanished for values near zero. This permits estimate from the data κ, η_0, and f_A (Efron, 2004). We applied a Bayes local false discovery rate (fdr) statistic (Efron, 2004) to asses the significance of the edges. The posterior probability that a specific edge exists given $\tilde{r}$ equals

$$IP(non\_null\ edge|\tilde{r}) = 1 - fdr(\tilde{r}) = 1 - \frac{\eta_0 f(\tilde{r};\kappa)}{f_{\tilde{r}}}$$

We considered an edge to be significant when the probability of an edge was > 0.8 (Efron, 2005).

Maximum relatedness subnetwork

To extract the essential associations within clusters for the resulting GGM network, we used the maximum relatedness subnetwork (MRS) approach (Lee et al., 2010). MRS algorithm assigns to each node i, a directed link to node j with which node i has its maximum and minimum partial correlation values. One node could have more than one directed link in the case of the multiple nodes with the same maximum partial correlation value. GGMs are undirected graphical models; using the MRS we were able to assign directionality to the edges and provide more information about the asymmetry of the node pairs. Node's incoming degree in MRS let us to identify the nodes with more importance in the GGM.

Metrics associated to network topologies have been calculated, those are average node connectivity, average path length, diameter, cumulative degree distribution, clustering coefficient and modularity. ForceAtlas 2 from Gephi software (Bastian et al., 2009) was used to layout the resulting network. All analyses involved in the data preparation and network analysis were performed using the R language and environment (Team, 2010). GGMs were calculated using an in-house modified version of the GeneNet package (Schäfer et al., 2001) and MRS was de novo implemented in R. GGMs were represented using Gephi.

Unknown clusters in sequenced genomes

In order to relate clusters of *unknowns* with the hypothetical proteins found in sequenced genomes we performed a BLAST+ tblastn search (Camacho et al., 2009) using as query all the *unknown* clusters against all microbial genomes available on Genbank release 190.0 (June 2012). We selected all hits that reported a hypothetical protein with an *e-value*$\leq 1e-05$.

GO and KO mapping

We mapped the PFAM families included in the MRS to their GO homologs using the mapping files derived from InterPro (Hunter et al., 2009) available in Gene Ontology (Harris et al., 2004). We performed a KEGG ORTHOLOGY (KO) mapping using the domain information found in KEGG GENES Release 58.1 (June 2011) (Kanehisa et al., 2008). Once we obtained the KO terms we did a pathway reconstruction using MinPath 1.2 (Ye et al., 2009)

Nitrosopumilus maritimus SCM1 subnetwork extraction

We performed a functional annotation of*Nitrosopumilus maritimus* SM1 (NC_010085.1) genome with HMMER 3.0 and PFAM 26.0. Then we selected all the genes with a hit PFAM domain associated and we mapped them to the existing PFAM domains present in the complete network. With this information we created a *Nitrosopumilus maritimus* centered subnetwork.

Metatranscriptome crosschecking of AMO neighborhood candidates

We downloaded from CAMERA all cDNAs for the metatranscriptomes CAM_P_0000692 (Oregon MT) and CAM_PROJ_Sapelo2008 (Sapelo MT) to find expression evidences of the nodes found in the AMO network neighborhood. For each neighbor node, we used the *Nitrosopumilus* proteins that contained the PFAM domain to perform a BLAST+ tblastn search in the metatranscriptomes. We reported the raw reads count for the hit with an *e-value*$\leq 1e-05$.

8.3 Results and discussion

Pre-network processing: Annotation, clustering and filtering.

From the original dataset containing 7,523,471 reads, HMMER was able to annotate 5,653,491 reads resulting in 15,528,086 hits. The *unknown* fraction achieved 1,869,980 reads and after performing a 6-frame translation, we kept a total of 8,884,278 translated sequences larger than 60 aminoacids. The hierarchical clustering at 90% reported 7,681,220 clusters and 6,689,553 clusters at 60%. We removed 5,759,646 singletons resulting in a total of 929,907 clusters. In the final data set, 9,190 of the reads were 16S rDNA and after clustering at 97% identity, ended in 7119 reference 16S rDNA. Sampling site filtering steps resulted in 6,903 different PFAM families, 9,925 *unknowns* clusters and 347 16S rDNA.

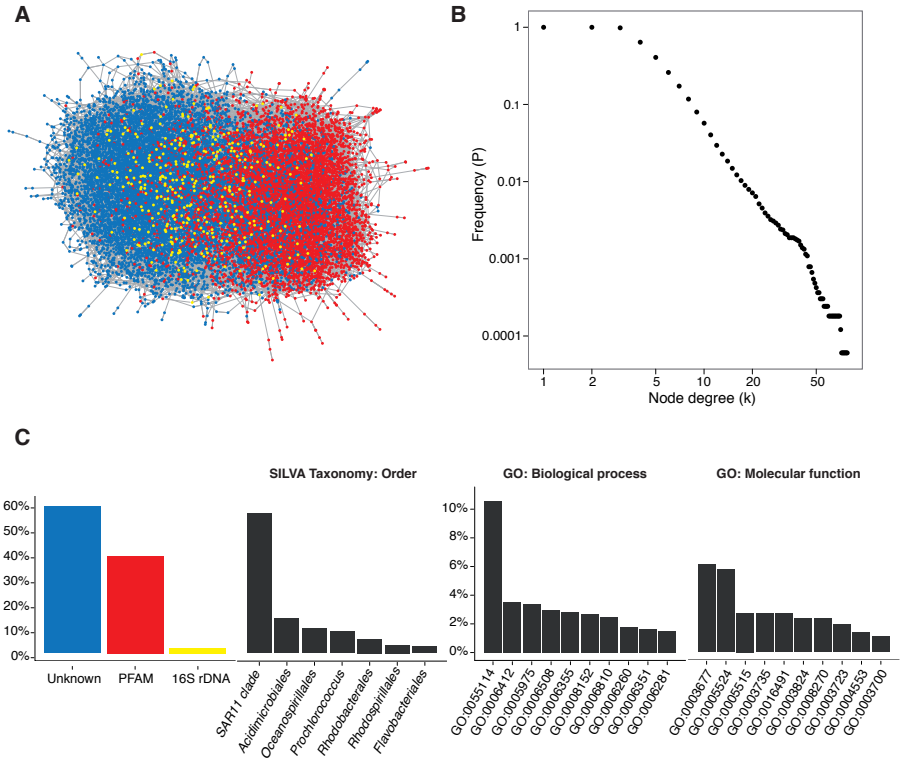

Figure 8.1: Global Ocean survey Maximum Relatedness Subnetwork (A) capturing 32672 associations among 16520 nodes. Each node represents an *unknown unknowns* cluster (blue), as 16S rDNA cluster (yellow) or a PFAM family (red). MRS node degree distribution (B) fit a power law distribution ($\alpha = 3.5, x_{min} = 12$ and $p > 0.1$) although as most biological systems is not a totally scale-free network. In C) there are representes the proportion of *unknown unknowns* clusters, 16S rDNA clusters and PFAM families and the taxonomic and functional composition of the MRS. GO terms for Biological Process are GO:0055114 oxidation-reduction process; GO:0006412 translation; GO:0005975 carbohydrate metabolic process; GO:0006508 proteolysis; G O:0006355 regulation of transcription, DNA-dependent; GO:0008152 metabolic process; GO:0006810 transport; GO:0006260 DNA replication; GO:0006351 transcription, DNA-dependent; GO:0006281 DNA repair. GO terms for Molecular function are GO:0003677 DNA binding; GO:0005524 ATP binding; GO:0005515 protein binding; GO:0003735 structural constituent of ribosome; GO:0016491 oxidoreductase activity; GO:0003824 catalytic activity; GO:0008270 zinc ion binding; GO:0003723 RNA binding; GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds; GO:0003700 sequence-specific DNA binding transcription factor activity.

GGM construction

Our initial matrix suffered from the "small n, large p". The matrix arrangement had 17,175 columns and 53 rows. The optimal shrinkage intensity (λ) estimated from the data was 0.4171. From the maximum number of 294,980,625 partial correlations we kept 552,099 with a $posterior probability > 0.8$; We reduced the complexity of the resulting network applying the MRS algorithm; the final subnetwork has 16,520 nodes and 32,672 edges. For the nodes, 6,459 represented PFAM families (47% of PFAM families), 9,714 clusters of *unknowns* and 347 nodes were 16S rRNA gene clusters (Figure 8.1A). As shown in Figure 8.1A, PFAM and clusters of *unknowns* were clearly divided in two groups on the MRS.

We calculated the networks metrics for the complete network (CN) and for the MRS; in both cases we interpreted the networks as undirected. The CN had an average degree of 66.84 while the MRS 1.98. Figure 8.1B represents the cumulative degree distribution of the MRS. As shown in Figure 8.1 the MRS seemed to follow a scale-free degree distribution (that is, a few nodes highly connected while the vast majority only have few connections). This network topology is typically found in biological systems (Bork et al., 2004) and in co-ocurrence domain networks (Wuchty, 2001). For a scale-free network the degree distribution has to follow a power law distribution, and we confirmed that the dataset fit in ($\alpha = 3.5$, $x_{min} = 12$ and $p > 0.1$). We also run a likelihood ratio test (LRT) in order to find whether other distributions could fit better the dataset. We found that the exponential distribution fit our data equally good as the power law (Clauset et al., 2007). Most biological networks are not totally scale-free but certain scale-free features such as small world and centrality properties were present (Khanin & Wit, 2006; Amaral et al., 2000).

The average path length (distance between all pair of nodes) was 2.9 for the CN and 7.4 for the MRS. The diameter of the CN was 8 while MRS had a diameter of 15 nodes. Clustering coefficient quantified how well connected are the neighbors of a node in a graph, and the CN had a clustering coefficient of 0.17 and 0.016 for MRS. Measuring the modularity index (Q) for the MRS, $Q = 0.573$, we detected 20 modules with a resolution of 2.5 (Fortunato & Barthelemy, 2007) considering edge weights in the calculations (Figure 8.2). We created subnetworks for the modules that accumulated more vertices and we extracted for the most abundant 16S rRNA their first order neighbors subnetwork as well.

GO and KEGG functional annotation

From the 6903 PFAM domains included in the network we mapped 2476 to a GO term. As a result of the many-to-many relationship between PFAM

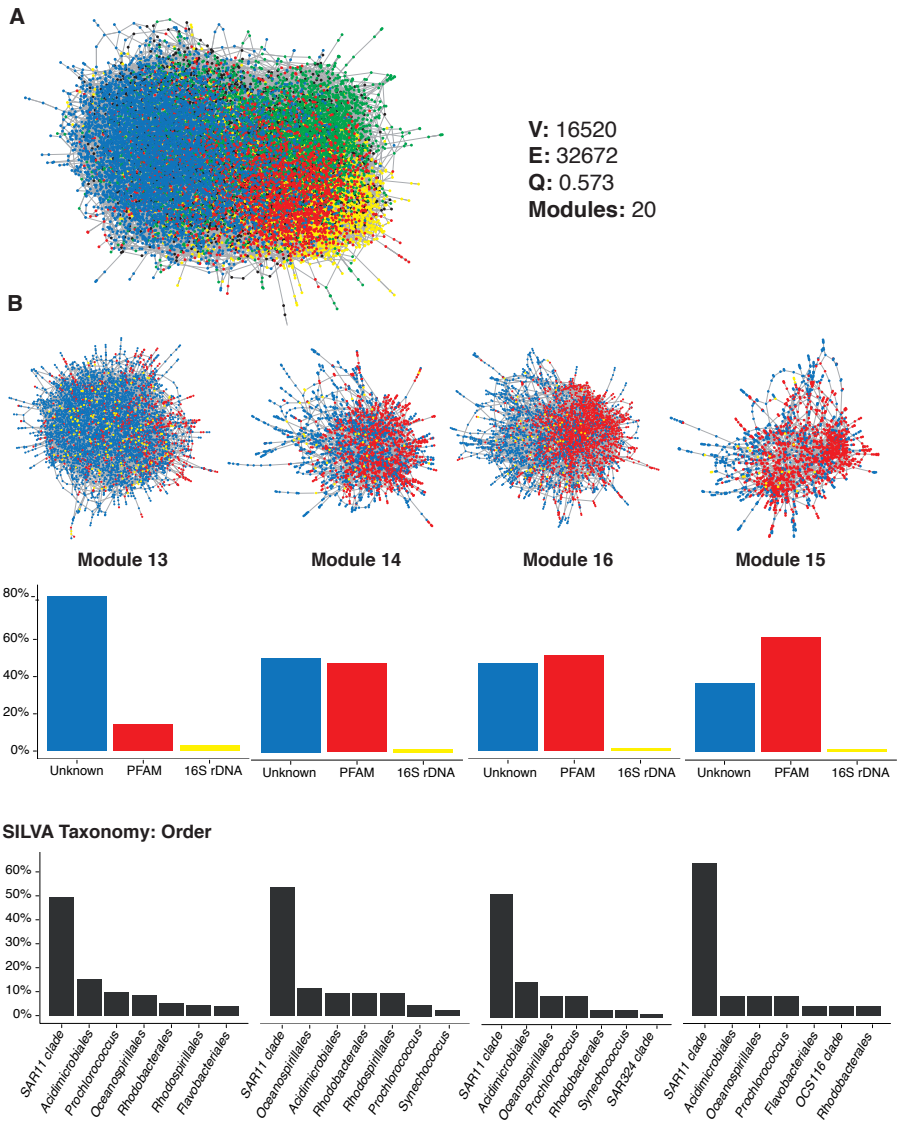

Figure 8.2: MRS modules detected by community detection methods. The four (A) more abundant modules are colored: Module 13 (blue), Module 14 (red), Module 15 (yellow) and Module 16 (green). In B) there are represented the proportion of *unknown unknowns* clusters, 16S rDNA clusters and PFAM families and the taxonomic composition for each module of the MRS.

and GO terms we got 5699 potential GO terms, of which 1516 were unique. The three most abundant GO terms (Figure 8.1C) in Biological Process domain were GO:0055114, oxidation-reduction process; GO:0006412, translation and GO:0005975, carbohydrate metabolic process. And for Molecular Function domain were GO:0003677, DNA binding; GO:0005524, ATP binding and GO:0005515, protein binding.

The mapping from PFAM domains to KEGG ORTHOLOGY resulted in 20932 possible annotations, and the effect of many-to-many relationships was even higher than mapping GO terms. We used those identified KO terms to run a KEGG pathway reconstruction using the parsimony approach implemented by MinPath (Ye et al., 2009). Protein domains contained in the MRS leaded to the reconstruction of 236 pathways composed by 9527 KO terms (Table S2). At least, 80 pathways had more than the 90% involved protein families recovered from our dataset. In our MRS we were able to reconstruct 92 of the 148 related pathways (each pathway had more than the 50% of proteins involved annotated) to the KO term Metabolic pathways (ko01100).

Clusters of *unknowns* in sequenced genomes

When we queried the clusters of *unknowns* against the sequenced genomes using BLAST, we obtained 1559 hits with hypothetical protein as annotation and 4933 clusters of *unknowns* had a hit to an annotated protein. Nearly the 70% of the hits reported had an e-value < 1e-20 (Figure S1; Table S1). A bit more than one third of the clusters of *unknowns* (3222) did not match to the sequenced genomes currently available.

Two factors have to be considered for the annotated protein hits obtained. First, the use of strict search parameters to rule out false positives caused by similar strong biased composition between the query and the hit. Second, although PFAM contains an exhaustive number of protein families, it did not fully cover the whole universe of proteins. For instance, PFAM release 26 has a coverage of nearly the 80% of proteins found in UniProtKB (Punta et al., 2012).

MRS modules analysis

Module 13 (32%), Module 14 (16%), Module 15 (12%) and Module16 (24%) accumulated up to the 84% of the nodes in the MRS. Such modules showed clear differences in terms of PFAM domains, clusters of *unknowns* and 16S rRNA clusters. Most of the Module 13 nodes were classified as clusters of *unknowns* (83%) and just a small fraction of nodes were PFAM domains (14%). Conversely, Module 15 showed 60% of the nodes as PFAM domains. Module 14 and Module 16 had a similar number of clusters of *unknowns* and PFAM

domains. By far, the bacterial SAR11 clade was the most abundant 16S rRNA gene cluster found in each case.

To unveil differences in the functional content for each module we reconstructed the biological pathways independently. The protein domains included in the four modules yielded the reconstruction of 216 different pathways with 8519 KO terms mapped. While 5952 KO terms were shared between modules, 2567 terms were module specific (Figure S2, Table S3-S10).

Interistingly, Module 13 presented specific KO terms (338) related with signaling and interaction pathways, e.g., Neuroactive ligand-receptor interaction pathway had 42% of specific KO terms. Module 14 (650 specific terms), showed most of them involved in different metabolic pathways. Module 15 had the 20% of its KO terms (750) into the two-component regulatory systems. And Module 16, with 829 specific terms, had ABC transporters as the pathway with more mapped KO terms (13%).

MRS taxonomic composition

The four most abundant 16S rRNA clusters in the MRS were *SAR11 clade* (53%), *Acidimicrobiales* (13%), *Oceanospirillales* (9%) and *Prochlorococcus* (8%). For each of these 16S rRNA clusters we extracted the first order neighbour subnetwork (Figure 3). The node neighborhood for SAR11 had almost the same proportion of PFAM domains (41%) and clusters of *unknowns* (45%) whilst the other 16S rRNA clusters had most of the nodes classified as clusters of *unknowns*, being *Prochlorococcus* the most extreme case with 68% of the nodes as clusters of *unknowns* and only a 13% of them, classified as PFAM domains.

We explored in detail the clusters of the *unknowns* in the neighborhood of SAR11 clade and *Prochloroccocus* in genomes as hypothetical proteins. From the 420 hits reported by BLAST for SAR11 clade, 30% of them were found in genomes from that clade (Figure 8.4). The percentage of hits increased when looking at a higher taxonomic level, and 52% of the hits matched the Alphaproteobacteria class. In Prochlorococcus, the 60% of *unknown* neighbors annotated as hypothetical proteins were found in genomes of the same order (Figure 8.5).

Functional and structural relationships found in the GOS MRS

One of the advantages of combining Graphical Gaussian Models and a Protein Domain Co-occurrence Network was the possibility to capture the structural interactions of domains within a protein and the functional interactions between proteins. Here we present some examples of the possibilities that this

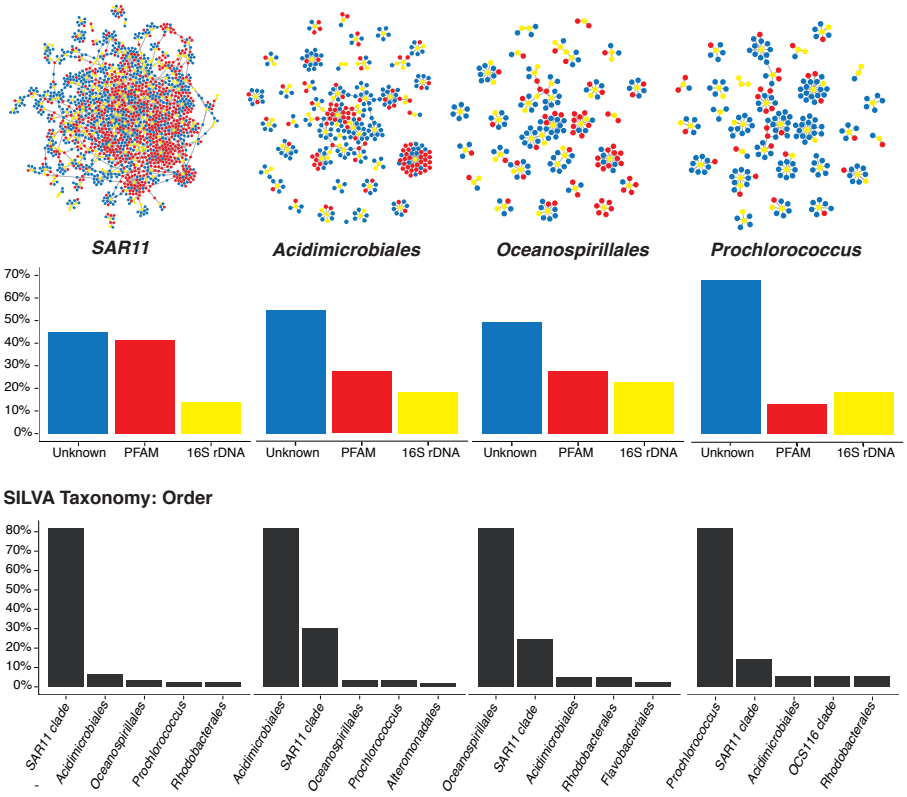

Figure 8.3: First order neighborhood subnetwork for the four more abundant taxonomic Order found in the MRS. Each node represents an *unknown unknowns* cluster (blue), as 16S rDNA cluster (yellow) or a PFAM family (red). Bar plots represents the proportion of *unknown unknowns* clusters, 16S rDNA clusters and PFAM families and the taxonomic and functional composition on each subnetwork.

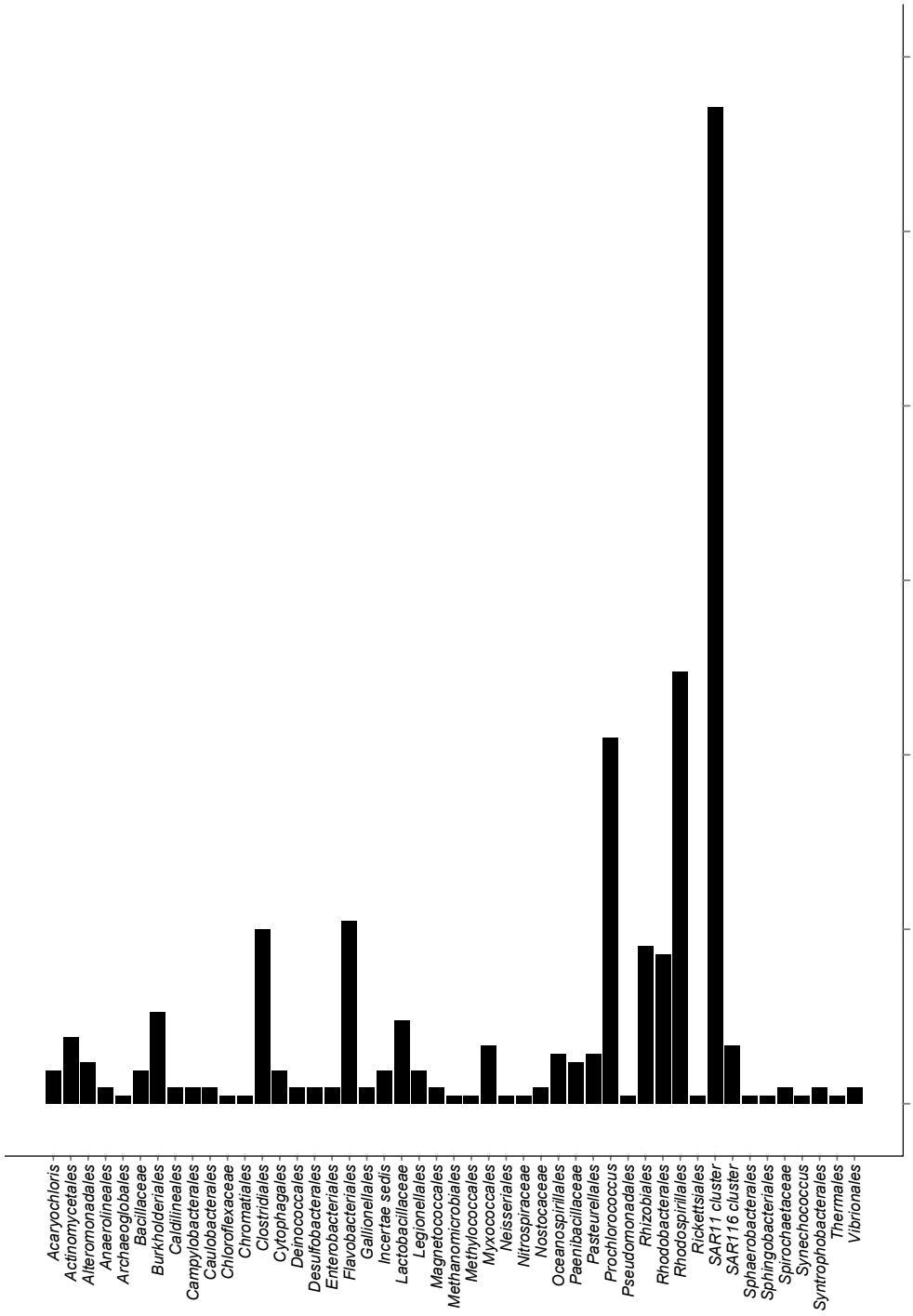

Figure 8.4: Taxonomic distribution of the hypothetical proteins identified on the *SAR11 clade* ego network.

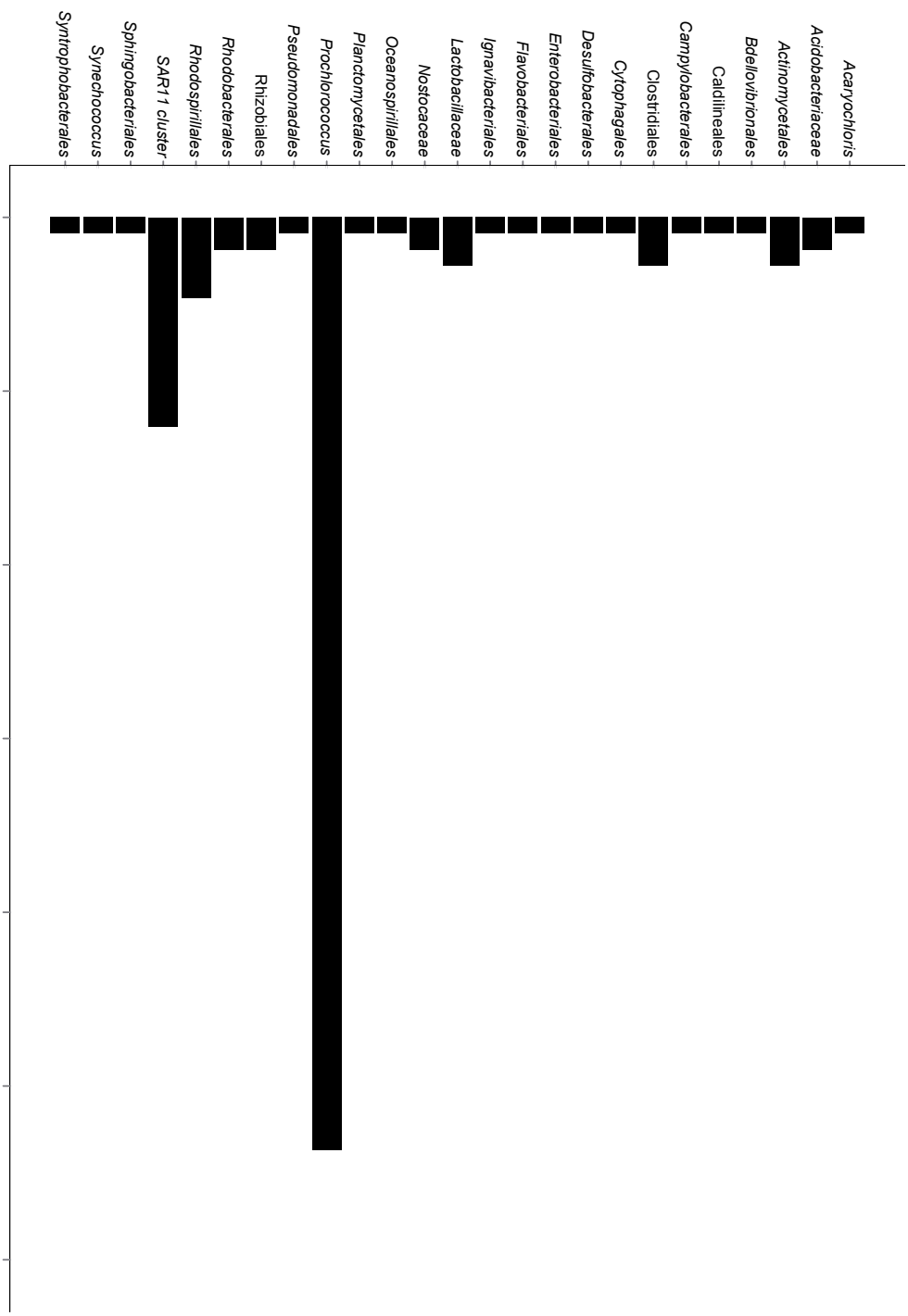

Figure 8.5: Taxonomic distribution of the hypothetical proteins identified on the *Prochlorococcus* ego network.

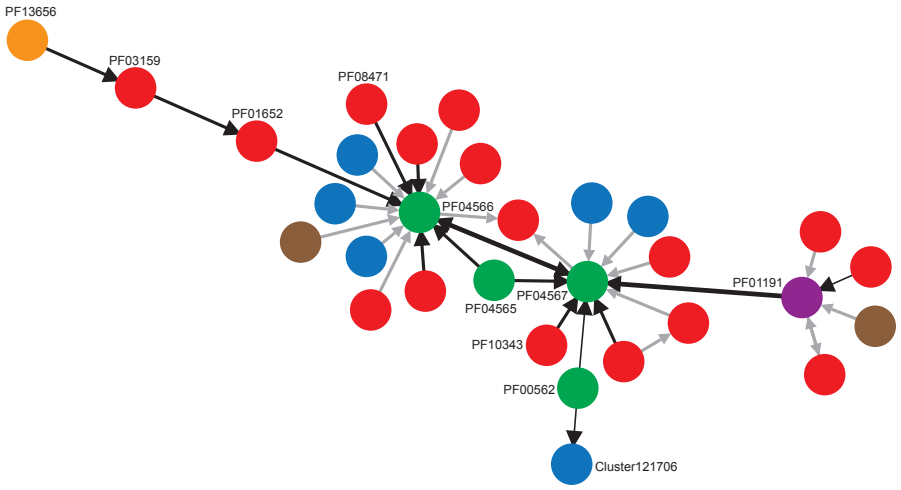

Figure 8.6: First order neighborhood subnetwork for the RNA polymerase II. Nodes color correspond to Rpb2 subunit domains (green), Rpb3-Rpb11 heterodimer (orange), Rpb5 subunit (lilac), PFAM families (red), *unknown unknowns* (blue) and *unknow unknows* with homology to a hypothetical protein found in a sequenced genome. Edges with a $\tilde{\rho}_{ij} < 0$ are colored in grey and edges with a $\tilde{\rho}_{ij} > 0$ in black. Edge thickness corresponds to the $\tilde{\rho}_{ij}$ strength.

approach could bring to broaden our knowledge in metagenomic studies and possibilities of unveiling new functions that remained unconnected.

RNA polymerase II subunit

Prokaryotes have a single type of RNA polymerase (RNAP), in Bacteria, RNAP is formed by the subunits α, β and ω whilst archaeal RNAP is composed by the Rpo1–Rpo13 subunits (Minakhin et al., 2001; Yee et al., 2000; Werner & Grohmann, 2011) . Bacterial β-subunit is the homolog to the RPB1 and RPB2 in Eukarya and Rpo1 and Rpo2 in Archaea. PFAM domain annotation uses the eukaryotic form, e.g. PF04565.11 is annotated as RNA polymerase Rpb2, domain 3, we will use the term specified in the PFAM annotation, in this case Rpb2, to refer to the prokaryotic RNAPs. Figure 8.6 shows the first order neighborhood subnetwork of the RNAP, in this network our approach has been able to recover associations that involve structural properties as functional aspects related with the transcriptional function.

Rpb2, formed by 8 domains, is the second largest subunit in RNAP and a key element in the core complex (Cramer et al., 2001); domains 3, 4, 5 and 6 from Rpb2 are present in the network and they are connected together (PF04565, PF04566, PF04567 and PF00562 respectively). PF04567, Rpb2 domain 5 has a link to PF01191, the subunit C from Rpb5, the homolog

counterpart of archael Rpo5 and is involved in DNA melting and early transcription (Miyao & Woychik, 1998; Yee et al., 2000; Grünberg et al., 2010).

PF13656 stands for the Rpb3-Rpb11 (the α subunit in bacteria) dimerization domain; Rpb3-Rpb11 heterodimer is involved in the initiation of the RNAP assembly((Benga et al., 2005; Kimura et al., 1997). Connected to the PF13656 we found PF03159, an XRN 5′-3′ exonuclease, that promotes the termination of RNAP transcription activity by the degradation of the downstream cleaved RNA (Kaneko et al., 2007; Moore & Proudfoot, 2009; Cramer et al., 2000).

In association with the termination related exonuclease and the Rpb2 domain 4, we found PF01652, and eukaryotic translation initiation factor member of the 4E family. Homologs of this translation factor are also found in Prokaryotes (Kyrpides & Woese, 1998).

In the neighborhood of Rpb2 domain 4, there is a ribonucleotide reductase domain, PF08471, involved in the reduction of ribonucleotides to deoxyribonucleotides for DNA synthesis and repair (Borovok et al., 2002) as the one found in the bacterial transcription-repair coupling factor/RNA polymerase interaction (Westblade et al., 2010).

Based on network topological measures like node degree, we are able to unveil functional differences based in structural properties like those observed in Rpb2 domain 3, 4 and 5. Domain 4 and 5 have a node degree of 15 and 12, these domains correspond to the external 2 region of Rpb2, while domain 3 and 6 with a node degree of 2 are the domains found in the fork and hybrid binding, inner regions of Rpb2 (Cramer et al., 2001), therefore with less potential to interact with other proteins as node degree reflects.

KaiC domain: Circadian clocks and recombinational repair

KaiC is a protein family that binds to DNA. Two members of this family, KaiC and RadA/Sms, are involved in different functional processes. KaiC proteins are responsible of the circadian rhythms in *Cyanobacteria*, which regulates global changes in the expression patterns in the cell. KaiC behaves as its own transcriptional repressor by autophosphorylation and is implicated on ATP-dependent hexameric rings formation and binding to forked DNA substrates (Mori et al., 2002).KaiC protein from *Synechococcus* is composed by two RecA-like domains (Leipe et al., 2000). Whilst RadA/Sms, an ATP-dependent protease, is involved in recombination and recombinational repair (Beam et al., 2002). RadA/Sms middle region is related to the RecA strand exchange protein and the DnaB replicative DNA helicase (Leipe et al., 2000).

Figure 8.7A shows the first order neighbor network centered in the KaiC domain PF06745. Its neighborhood defines the differences and similarities between KaiC and RadA/Sms proteins. In the network, node PF00154 is the

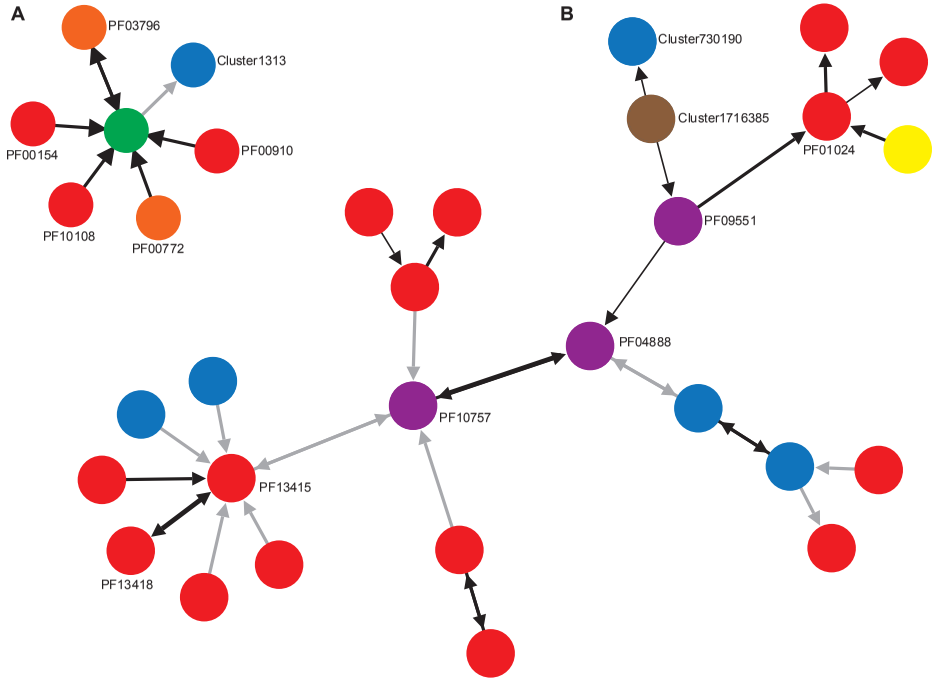

Figure 8.7: First order neighborhood subnetwork for the KaiC domain (A) and biofilm formation and regulation domains (B). Red colored nodes correspond to the PFAM families, blue colored nodes represents the *unknown unknowns*, 16S rDNA clusters are colored in yellow and and *unknow unknows* with homology to a hypothetical protein found in a sequenced genome in brown. In A, KaiC domain is colored in green, C- and N-terminal domain of a DNAB-like helicase are colored in orange. In B, lilac colored nodes correspond to the PFAM families involved in biofilm formation and regulation. Edges with a $\tilde{\rho}_{ij} < 0$ are colored in grey and edges with a $\tilde{\rho}_{ij} > 0$ in black. Edge thickness corresponds to the $\tilde{\rho}_{ij}$ strength.

domain for the bacterial recA bacterial DNA recombination protein. In both proteins, a RecA-like domain is present. RadA/Sms recombinational repair function is carried out by the combination of PF03796, PF00772 and PF10108. PF03796 and PF00772 are the C- and N-terminal domain of a DNAB-like helicase; and PF10108 is an exonuclease related to the exonuclease domain of PolB, a DNA polymerase that participates in DNA repair and replication.

Biofilm formation and regulation

Biofilms are an aggregate of microorganisms in which cells adhere to each other on a surface embedded within a self-produced matrix of extracellular polymeric substance composed basically from proteins and polysaccharides. This matrix acts as protection for the cells and facilitates communication through biochemical signals. In the subnetwork shown in Figure 8.7B, PF10757 is the domain of the regulator in biofilm formation, YbaJ. This protein regulates biofilm formation and also has an important role in the regulation of motility in the biofilm. YbaJ stimulates conjugation, aggregation and decreases motility, resulting in an increase of the biofilm (Barrios et al., 2006). Related with cell communication we found PF04888, a secretion system effector C (SseC) like family. SseC family includes the bacterial secreted proteins PopB, PepB, YopB and EspD that are involved in pore formation and type III secretion system translocon. Type III secretion system are part of the direct intercellular transport of molecules complex (Beloin et al., 2008; Margolis et al., 2010; Christopher et al., 2010) . Associated to SseC we found PF09551, a stage II sporulation protein, where SseC family proteins could have a function in the signaling mediated by a channel that links the mother cell to the forespore (Camp & Losick, 2009, 2008) . In the network we found PF01024, a colicin pore-forming domain; colicins are released into the environment to reduce competition. Negatively associated we found PF13418 and PF13415, a galactose oxidase domain that catalyzes the chemical reaction $D-galactose + O_2 \rightleftharpoons D-galacto-hexodialdose + H_2O_2$. D-galactose is a key component for the synthesis of the exopolysaccharide, critical molecule for biofilm formation. (Ma et al., 2007; Chai et al., 2012; Ryder et al., 2007) .

Ammonia oxidization in *Nitrosopumilus maritimus SCM1*

In the bacterial ammonia oxidation, the ammonia monoxygenase enzyme produces hydroxylamine that is further oxidized by the hydroxylamine oxireductase (HAO) complex. When hydroxylamine is oxidized it supplies electrons to the ammonia oxidase and to the bacterial typical iron-based electron transfer electron transport chain. The AOA have an alternative pathway without HAO and cytochrome c proteins. In Archaea, the electron transfer mechanism

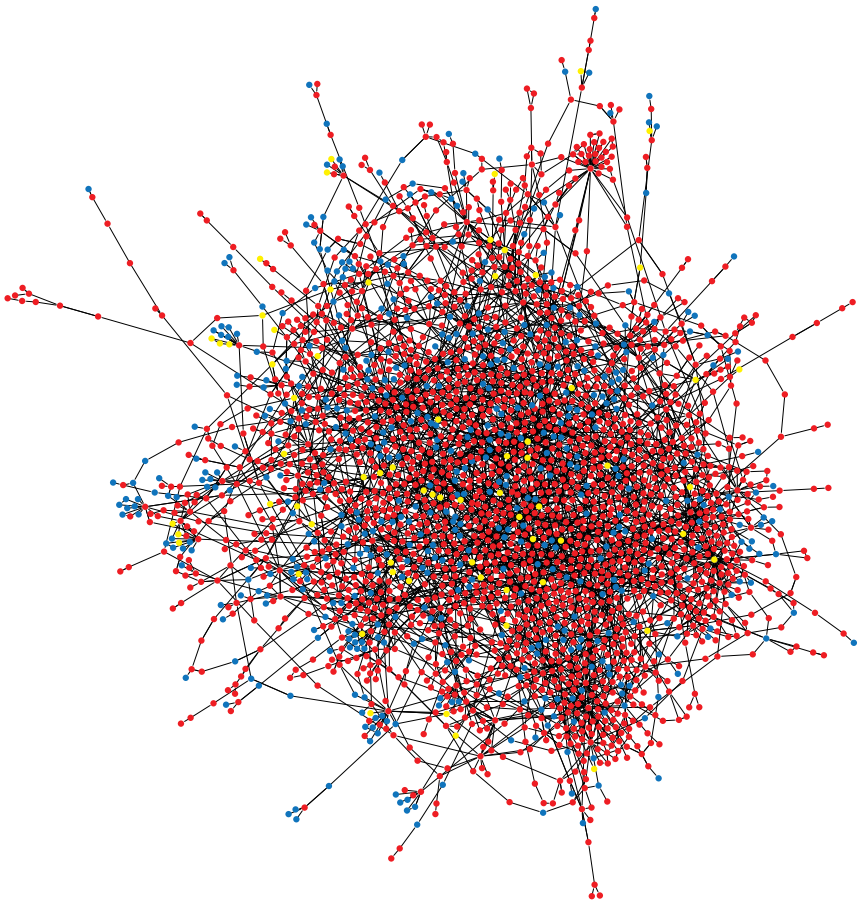

Figure 8.8: *Nitrosopumilus maritimus SCM1* genome centred subnetwork. Each node represents an *unknown unknowns* cluster (blue), as 16S rDNA cluster (yellow) or a PFAM family (red).

seems to be mediated by the numerous copper-containing proteins contained in their genomes. In the ocean, members of the Marine AOA are one of the most abundant microorganisms (Karner et al., 2001; Agogué et al., 2008) and are one of the most significant contributors to carbon and nitrogen cycling.

To find the association between the AMO subunits and their networks we used the *Nitrosopumilus maritimus* genome subnetwork. We were able to recover 2277 PFAM domains that belonged to 1342 *Nitrospumilus maritimus* proteins, nearly the 75% of the whole proteome.

Figure 8.9 represents the first order neighbor network for the archaeal *amoA* and amoB subunits. It worthy to note, that the 57% of neighbor nodes to AMO subunits were PFAM domains found in *Nitrosopumilus maritimus* SCM1 genome. This proportion increases up to the 92% if we remove the negatively associated clusters and the two PFAM domains with the lowest partial correlation strength that were not identified in the genome. As has been shown in the previous examples, those 9 positively associated nodes to the AMO subunit domains, most probably will have a role in the archaeal ammonia oxidation.

In the neighborhood of *amoA* we found the domain family PF03911.11 that corresponds to a Sec61β family. Although this domain is not present in the annotation of *Nitrosopumilus maritimus* SCM1, we identified it in an ORF located at chromosomic positions 563213..563049. This family consists of homologues of Sec61beta, a component of the Sec61/SecYEG protein secretory system. These translocases are responsible for decoding the topogenic sequences within membrane proteins that direct membrane protein insertion and orientation (Grassly & Fraser, 2008; Kalies et al., 1998; Park & Rapoport, 2011; Kinch & Saier Jr, 2002). Next domain in the *amoA* neighborhood was PF01950.11, a Fructose 1,6-bisphosphatase domain encoded by the gene Nmar_1035. Fructose-1,6-bisphosphatase is an essential regulatory enzyme in the gluconeogenesis. (Nishimasu et al., 2004). The node PF01862, a Pyruvoyl-dependent arginine decarboxylase domain, is product of the gene Nmar_1180. This domain produces agmatine after arginine decarboxylation (Graham et al., 2002). Agmatine is an essential intermediate in polyamine biosynthesis (Fukuda et al., 2008) and in translation((Osawa et al., 2011). Agmatinase (Nmar_0925) hydrolyses agmatine to putrescine, the precursor for the biosynthesis of higher polyamines like spermidine and spermine (Ahn et al., 2004). Spermine/spermidine plays an important role as a substrate for the hypusine (Nmar_0648) modification of the eukaryotic translation initiation factor 5A (eIF-5A), encoded by the gene Nmar_0529. eIF-5A is involved in cell growth and protein synthesis (Peat et al., 1998; Zanelli & Valentini, 2007; Wagner & Klug, 2007; Kim et al., 1998; Chattopadhyay et al., 2008; Eichler & Adams, 2005; Jansson et al., 2000; Wolff et al., 2007).

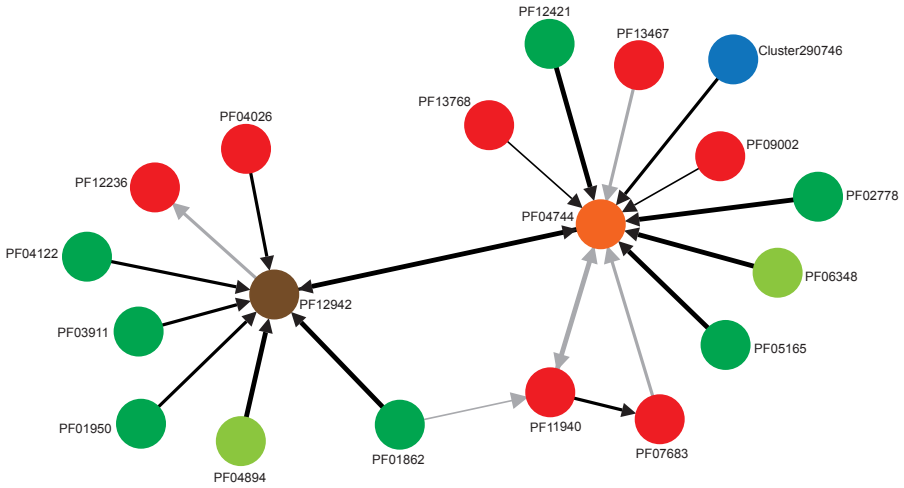

Figure 8.9: First order neighborhood subnetwork for the AMO subunits A (brown) and B (orange). Red colored nodes correspond to the PFAM families, blue colored nodes represents the *unknown unknowns*. Dark green nodes represent PFAM families found in proteins coded by genes in *Nitrosopumilus maritimus* SCM1 genome. Light green nodes are DUF domains found in *Nitrosopumilus maritimus* SCM1 genome. Edges with a $\tilde{\rho}_{ij} < 0$ are colored in grey and edges with a $\tilde{\rho}_{ij} > 0$ in black. Edge thickness corresponds to the $\tilde{\rho}_{ij}$ strength.

In the amoB neighborhood we found, PF05165, a GTP cyclohydrolase III product of the gene Nmar_1614, that produces a formylaminopyrimidine from the hydrolysis of GTP that is the starting point for riboflavin biosynthesis in many archaea; precursor for FMN and FAD biosynthesis, two redox cofactors involved in many steps in the metabolism (Graham et al., 2002; Shiemke et al., 2004; Mashhadi et al., 2008).

Another node associated to the amoB subunit was the domain PF02778.9, a tRNA intron endonuclease, product of gene Nmar_0450 and Nmar_1039. It is interesting to note the relationship between AMO subunits and domains related to tRNA processing (Heinemann et al., 2010). On the one hand, it is related to the tRNA splicing, perhaps as a response to stress factors it could trigger the translation of codon-biased proteins (Chan et al., 2012); on the other hand is related with agmatine. Agmatine is also important for the agmatinylation of tRNA, and essential mechanism for the correct AUA codon decoding (Osawa et al., 2011; Mandal et al., 2010).

Finally, from the neighborhood of AMO, we selected three candidates to be potentially related structurally or functionally to the ammonia oxidation in some way. The proteins in *Nitrosopumilus* with PF06348 and PF12421 domains, are good candidates to be directly involved into ammonia oxidation or electronic transfer by their special structure. PF06348 is a domain of *un-*

known function (DUF1059) and is the protein product of the gene *Nmar_0235*. This protein seems to be specific to the *Nitrosopumilales* genomes. We analyzed more in detail *Nmar_0235* using DELTA-BLAST (Boratyn et al., 2012) to find distant homologs; and we found that it shares homology with predicted small metal-binding proteins (COG5466). As a result of the higher concentrations of copper in seawater (Hatzenpichler, 2012) and its association with the amoB subunit, this small protein could have a role in cellular copper management, similar to the one observed in the protein SmbP found in *Nitrosomonas europaea* (Barney et al., 2004). Domain PF04894, an archaeal protein of *unknown* function (DUF650), is the product of the gene Nmar_0975 and is conserved along the entire archaeal domain. Its possible implication on archaeal ammonia oxidation remains unclear, we were unable to find any clue to a possible function. The last candidate was PF12421.3, a Fibronectin type III protein domain (FnIII). We found it present in three different genes, *Nmar_1073*, *Nmar_1047* and *Nmar_1040*. *Nmar_1073* is a giant gene (28758 nucleotides) that encodes a protein with 73 FnIII repeats. Nmar_1047 contains a NHL repeat in the N-terminal end followed by three FnIII repeats. While *Nmar_1040* encoded a protein formed by a signal peptide spanning positions 1 to 34, followed by three FnIII repeats and with a multicopper oxidase-like domain at the C-terminal end. Those three genes, after a screening by BLAST against all microbial genomic proteins, seem to be specific to the *Nitrosopumilales* genomes.

FnIII domains in prokaryotic proteins are associated with a wide range of functions. They are part of carbohydrate processing enzymes such as cellulases and chitinases; are involved in adhesion and in cell-to-cell interaction functions when are located in cell surface and could assist large enzyme complexes to remain soluble. FnIII domains assist in the adhesion of the protein to polysaccharides or cell receptors keeping the proper conformation at the binding or catalytic sites of the enzyme (Little et al., 1994; Reva & Tümmler, 2008; Alahuhta et al., 2010).

Nmar_1040, as node associated to the amoB subunit, presents a highly interesting combination of domains. The multicopper oxidase-like domain at the N-terminal end has homology to a blue (type 1) copper protein and could be involved in electron transfer. Despite we have found a low sequence similarity (PFAM *e-value* $= 0.03$; Gene3D *e-value* $= 9.4e-21$; RPSBlast *e-value* $= 3.29e-07$) we can speculate about the existence of a potential copper-binding site. Similar to the structure observed in *Nmar_1040*, it has been previously observed in Geobacter with the *ompB* gene product. *OmpB* participates in the Fe(III) oxide reduction, and contains a Fe(III)-binding site, a fibronectin type III domain and a multicopper domain (Mehta et al., 2006; Holmes et al., 2008; Schäfer et al., 1999). The structural analogy between *Nmar_1040* and

ompB, and its association to amoB subunit strongly suggests *Nmar_1040* is a key component in the ammonia oxidation with a potential relevant role that has not been unveiled yet.

Interestingly, we found evidences of expression for the three candidates on the two marine metatranscriptomes we used for crosschecking. We found 58 different transcripts in the Oregon MT and 6 on Sapelo MT for *Nmar_1040*; 23 different reads in Oregon MT and 2 in Sapelo MT for *Nmar_1073*; and 14 in Oregon MT and 1 in Sapelo MT for Nmar_1043.

8.4 Discussion

We have shown through different examples that the approach resulted in a tool helpful for the exploration of the information on complex systems contained in metagenomes, and for a better understanding of the biological processes operating in the ocean. One of the strengths of the approach is the inclusion in the network of the batch of information related to the clusters of *unknowns* (both Domains of Unknown Function, and full *unknowns*). With the inclusion of the *unknown unknowns* in the network, we investigated not only the *known* associations, but also the *unknown* universe, generating hypothesis on potential candidates that may have been playing a key role in biological processes but have remained hidden until now.

The use of Graphical Gaussian Models has more advantages than the relevance network methods, because measure the degree of dependence between protein domains and rule out the effects of a third variable when looking for direct association (Schäfer & Strimmer, 2005b). With the MRS, it is possible to extract a subnetwork containing the strongest associations and assign a direction to the edges showing the sense of the dependence between nodes.

Finally, we generated a specific *Nitrosopumilus maritimus* subnetwork and map the protein families related to the AMO subunit to specific genes in *Nitrosopumilus maritimus* SCM1 genome. Within the nodes associated to the AMO subnetwork we proposed a plausible explanation on how a member of the Sec61β family translocates AMO subunits into the membrane and how AMO is linked to the tRNA processing showing its implications on cell growth and regulation. We also unveiled two potential candidates, which by their special structural conformation, could be key components for the direct ammonia oxidation or components of the electron transfer system. Further exploration of these proteins by experimental assays could provide relevant information on how ammonia oxidation works in archaea.

Acknowledgements

This work was supported by grant EU-COST Action number ES1103: Microbial Ecology & The Earth System: Collaborating for Insight and Success with the new generation of sequencing tools (CISME) and grant CONSOLIDER-INGENIO2010 GRACCIE CSD2007-00067. This work has been calculated at HPC facilities at MPIMM in Bremen.

Appendix [2]

[2] See more Supplementary Information in http://nodens.ceab.csic.es/ecoevo/ch8/

General overview

9
Concluding Remarks

For many years microbiology, ecology and evolution have not been walking hand in hand together. Evolutionary microbiologists were traditionally more interested in studying genetic patterns in terms of exploring the links between genes and genomes and the mechanisms generating diversity and the selection operating processes. Meanwhile, microbial ecologists were more focused on the study of the relationships between microbes and their environment including other organisms and how these relationships modulate the global biogeochemical cycling. Luckily, microbial ecology is a very dynamic research field and well aware on the limitations to study microorganims has quickly integrated and adapted to microbial systems the improvements not only in methodology but also in concepts from other research areas such as macroecology. However, although microbial ecology and microbial evolution have phylogenies as the nexus point, how they approached to the topic was very different in each field. While evolutionists were using phylogenies to understand the past (identification of important evolutionary events), microbial ecologists were using it to understand the present (current biodiversity).

As stated before, microbial ecology is always redefining itself, and rapidly adopted the methods developed from the recent merging between community ecology and phylogenetics. Just a few years ago community ecologists started to be interested in bringing together evolutionary history and phylogenetic relationships of organisms to answer question regarding community assembly and diversity (Webb et al., 2002; Webb, 2000; Cavender-Bares et al., 2009). This new approach links short-term local processes and global processes that occur over deep evolutionary time scales. Phylogenetic community ecology studies unveil which processes are driving the community assembly and show the importance of evolution in the assembly process. The use of phylogenies has

revealed how important are community interactions for speciation, adaptation and extinction.

Microbial ecologists have been, and still are, eager to explore the huge diversity of microorganisms and have carried out an extensive sequencing effort for the 16S rRNA and several other markers genes from a large range of natural and artificial environments. As a result of such effort, microbial ecology provides a large amount of molecular data ready to be used for phylogenetic analyses. The combination of increasing computer power and larges phylogenetic datasets, have brought microbial ecology to the perfect scenario for the community phylogenetic approach. Thus, although this is a very recent area of research, microbial community phylogenetics has produced lot of fresh knowledge in a very short time (Jones & Hallin, 2010; Auguet et al., 2010; Barberán & Casamayor, 2010; Barberán et al., 2011).

In this PhD thesis I have applied these approaches to compare bacterial ammonia oxidizers (AOB) against their archaeal counterparts (AOA) using the ammonia monooxygenase subunit A gene (amoA). The results of the analyses revealed for the first time a global picture of the phylogenetic community structure of ammonia-oxidizing assemblages. Our study unveiled larger phylogenetic richness in AOA with more dissimilar communities and clear monophyletic groups for the different habitats. The rates of diversification in AOA were higher than in AOB and the archaeal diversification dynamics showed an unusual feature, with an initial diversification process followed by a long period of stasis and a final burst of diversification. The variations observed between AOB and AOA in terms of community structure, phylogenetic diversity, diversification patterns, and habitat dispersion were unexpected just a very few years ago, and the community phylogenetics approach has nicely captured these differences.

Understand the diversification processes observed in AOA and their successful performance under a myriad of different environmental conditions such as low pH, different ammonia concentrations, high hydrostatic pressures, high light exposure, low oxygen availability among others, needs however of a deeper insight adding the evolutionary processes. The ecological and genetic diversity observed in AOA can be explained at two different levels. First, processes related with a global cell adaptation to the different environments, probably because of a diverse gene content in the genomes of the different AOA groups. To link changes in genomic content with environmental conditions, additional AOA genomes are needed to carry out proper comparative genomic analyses. Currently, with only a few genomes available, some differences have been already shown (Stahl & de la Torre, 2012), i.e., only two small plastocyanin-like proteins are shared by all AOA. And second, processes related with molecular adaptations of the enzyme *amoA* to

efficiently carry out ammonia oxidation activity under very diverse environmental conditions, probably mediated by nucleotide substitutions. With the huge molecular data sets available for AOA it is feasible to explore in detail the molecular level and understand how small changes are triggering the diversification processes and ecological adaptations observed at a global scale. This has been nicely shown by, Bielawski et al. (2004) suggesting that the observed vertical distribution of photochemically divergent marine proteorhodopsins could be tuned by Darwinian selection. In this work they identified which lineages and sites potentially were under the effects of Darwinian selection and how these changes modulated a global effect.

The work by Bielawski et al. (2004), applied codon-based models (Li et al., 1985b; Nei & Gojobori, 1986; Goldman & Yang, 1994; Suzuki & Gojobori, 1999; Yang & Nielsen, 2000; Zhang et al., 2005) to analyze signatures of Darwinian selection. These methods have been extensively used in eukaryotes, viruses, sometimes with pathogenic bacteria and in a few cases in microbial ecology. Although the methods used by Bielawski et al. (2004) are based on the estimation of dN/dS and avoided to make any assumptions regarding the demographic history of the population, unlike other "neutrality tests" (Tajima, 1989, 1996; Fu & Li, 1993a,b; Deng & Fu, 1996; Fu, 1997; Misawa & Tajima, 1997; Fay & Wu, 2000), in prokaryotes there are some limitations to be aware. First, dN/dS is inappropriate when comparing either very distantly-related strains (dS saturated with multiple substitutions), or very closely-related strains, within which dN/dS is inflated by segregating nonsynonymous polymorphism (Rocha et al., 2006; Kryazhimskiy & Plotkin, 2008). And second, phylogenetic discordance caused by recombination affect likelihood methods for quantifying selection pressure on codon alignments and may suffer from high rates of false positives (Anisimova et al., 2003; Shriner et al., 2003).

For codon-based analyses, PAML software (Yang, 2007) has been the most widely used approach with more than 4700 citations. However, it has some limitations for the treatment of massive datasets, as those that microbial ecologists generated, and along this PhD thesis HYPHY (Pond et al., 2005)was applied. HYPHY can perform all models implemented in PAML and has its own high-level programming language (HyPhyBatch Language, HBL) that allows users to implement and distribute new methods of sequence analysis or to modify existing settings. And also has support for parallel computing, a very convenient feature when the analyzed data sets are huge as in this PhD thesis. HYPHY authors are constantly developing new cutting-edge methods and models that resulted very helpful for the work presented here.

Following Bielawski et al. (2004), individual changes at the level of nucleotides were translated to the global diversification patterns of archaeal ammonia oxidizers. Thus, this resulted in a step further from the results

obtained after applying community phylogenetics methods providing precise evolutionary information behind the phylogenetic patterns observed within an ecological context. We will gain the full picture once the results can be integrated in a comparative genomics framework.

Recently microbial ecologists are applying network analysis approaches derived from the science of complex systems, to analyze co-occurrence and correlation patterns observed in the environment to predict species interactions and develop ecosystem-wide dynamic models. These approaches have been initially applied to the large 16S rRNA gene datasets obtained from pyrosequencing to analyze species interactions. In this PhD thesis, the approach has been applied to explore functional genes interactions mostly related to the ammonia oxidizing process, and to confront one of the main challenges of metagenomics, the exploration of the *unknown* fraction of the metagenomic protein universe. Applying methods of *reverse engineering of regulatory networks* (Schäfer et al., 2001) the associations between the known and the unknown fraction were reconstructed offering a pioneering fresh view for microbial ecology. One especially relevant result obtained from this approach on AOA was the reconstruction of the association network of the different AMO subunits to the other proteins previously reported in the marine AOA *Nitrosopumilus*. The information recovered from metagenomics combined with available genomes fuels hypothesis for the particular and yet unknown biochemistry of ammonia oxidation in Archaea. The information contained in Chapter 8 provides candidates that could be involved in the unknown steps of this metabolic pathway giving specific gene names for further testing in lab experiments, and guiding future research to gain better understanding on the ecological, phylogenetic and evolutionary differences observed between AOA and AOB.

As a closure for the experience gained along my PhD, I am able to glimpse the next direction for microbial ecology. We are living in a time where every day, Next-Generation Sequencing (NGS) technologies provide more and more sequences, increasing the sequencing deep of any environment. The ideas and results shown in this PhD combined with the ideas from Whitaker & Banfield (2006) for the use of *population genomics* to analyze microbial communities and the concept of *reverse ecology* from Li et al. (2008b) to provide a new way to analyze microbial communities, will probably feed the newest revolution in molecular microbial ecology of the NGS era.

10
Conclusions

The general conclusions of this PhD dissertation are:

- AOA and AOB present different community structure in terms of phylogenetic richness and β-diversity.
- Diversification patterns showed a more constant cladogenesis through time for AOB whereas AOA apparently experienced two fast diversification events separated by a long steady-state episode. Diversification rates (γ statistic) for most of the habitats indicated $\gamma_{AOA} > \gamma_{AOB}$.
- The combination of codon-based models and community phylogenetic methods resulted in a valuable tool to understand the diversification process in environmental marker genes.
- The *amoA* gene showed evidences to be under purifying selection due functional constraints. However, strong evidences of episodic diversifying selection for individual sites and for lineages were also detected.
- *T*he amoA diversification pattern shows evidences of generation and maintenance of an evolutionary AOA seed bank driving the adaptive radiation into the different *amoA* phylogenetic clusters.
- *T*he amoA gene in the ocean shows different patterns of episodic diversifying selection to environmental conditions for *shallow* and *deep* ecotypes. The amoA from the OMZ is under strong functional constrains and did not show evidences of diversifying selection.
- The pioneering approach combining families of *unknowns* and Graphical Gaussian Models to analyze functional associations in metagenomes is a valuable tool for functional discovery.

Bibliography

Bibliography

ABE, T., KANAYA, S., UEHARA, H. & IKEMURA, T., 2009. A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses. *DNA Res*, **16**(5), 287–97.

ACKERLY, D., 2009. Phylogenetic Methods in Ecology. *Encyclopedia of Life Sciences (ELS)*.

AGOGUÉ, H., BRINK, M., DINASQUET, J. & HERNDL, G. J., 2008. Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. *Nature*, **456**(7223), 788–91.

AHN, H. J., KIM, K. H., LEE, J., HA, J.-Y. Y., LEE, H. H., KIM, D., YOON, H.-J. J., KWON, A.-R. R. & SUH, S. W., 2004. Crystal structure of agmatinase reveals structural conservation and inhibition mechanism of the ureohydrolase superfamily. *J Biol Chem*, **279**(48), 50505–13.

ALAHUHTA, M., XU, Q., BRUNECKY, R., ADNEY, W. S., DING, S.-Y. Y., HIMMEL, M. E. & LUNIN, V. V., 2010. Structure of a fibronectin type III-like module from Clostridium thermocellum. *Acta Crystallogr Sect F Struct Biol Cryst Commun*, **66**(Pt 8), 878–80.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J., 1990. Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

AMARAL, L. A. N., SCALA, A., BARTHÉLÉMY, M. & STANLEY, H. E., 2000. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, **97**(21), 11149.

ANISIMOVA, M., BIELAWSKI, J. P. & YANG, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular biology and evolution*, **19**(6), 950–958.

ANISIMOVA, M., NIELSEN, R. & YANG, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**(3), 1229–36.

ANISIMOVA, M. & YANG, Z., 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol*, **24**(5), 1219–28.

AUGUET, J.-C., BARBERAN, A. & CASAMAYOR, E. O., 2010. Global ecological patterns in uncultured Archaea. *ISME J*, **4**(2), 182–90.

AUGUET, J. C. & CASAMAYOR, E. O., 2008. A hotspot for cold crenarchaeota in the neuston of high mountain lakes. *Environmental microbiology*, **10**(4), 1080–1086.

AUGUET, J.-C. C., NOMOKONOVA, N., CAMARERO, L. & CASAMAYOR, E. O., 2011. Seasonal changes of freshwater ammonia-oxidizing archaeal assemblages and nitrogen species in oligotrophic alpine lakes. *Appl Environ Microbiol*, **77**(6), 1937–45.

AUGUET, J.-C. C., TRIADÓ-MARGARIT, X., NOMOKONOVA, N., CAMARERO, L. & CASAMAYOR, E. O., 2012. Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophic lake. *ISME J*.

BARBERÁN, A. & CASAMAYOR, E. O., 2010. Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. *Aquatic Microbial Ecology*, **59**, 1–10.

BARBERÁN, A., FERNÁNDEZ-GUERRA, A., AUGUET, J. C., GALAND, P. E. & CASAMAYOR, E. O., 2011. Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota. *Molecular Ecology*, **20**(9), 1988–1996.

BARNEY, B. M., LOBRUTTO, R. & WILSON, A., 2004. Characterization of a small metal binding protein from Nitrosomonas europaea. *Biochemistry*, **43**(35), 11206–11213.

BARRIOS, A. F. G., ZUO, R., REN, D. & WOOD, T. K., 2006. Hha, YbaJ, and OmpA regulate Escherichia coli K12 biofilm formation and conjugation plasmids abolish motility. *Biotechnology and bioengineering*, **93**(1), 188–200.

BASTIAN, M., HEYMANN, S. & JACOMY, M., 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks.

BEAM, C. E. . E., SAVESON, C. J. . J. & LOVETT, S. T. . T., 2002. Role for radA/sms in Recombination Intermediate Processing in Escherichia coli. *Journal of Bacteriology*, **184**(24), 6836–6844.

BELOIN, C., ROUX, A. & GHIGO, J. M., 2008. Escherichia coli biofilms. *Bacterial Biofilms*, 249–289.

BEMAN, J. M., POPP, B. N. & FRANCIS, C. A., 2008. Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. *ISME J*, **2**(4), 429–41.

BEMAN, J. M., SACHDEVA, R. & FUHRMAN, J. A., 2010. Population ecology of nitrifying Archaea and Bacteria in the Southern California Bight. *Environ Microbiol*, **12**(5), 1282–92.

BENGA, W. J., GRANDEMANGE, S., SHPAKOVSKI, G. V., SHEMATOROVA, E. K., KEDINGER, C. & VIGNERON, M., 2005. Distinct regions of RPB11 are required for heterodimerization with RPB3 in human and yeast RNA polymerase II. *Nucleic Acids Res*, **33**(11), 3582–90.

BIELAWSKI, J. P., DUNN, K. A., SABEHI, G. & BÉJÀ, O., 2004. Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc Natl Acad Sci U S A*, **101**(41), 14824–9.

BILLER, S. J., MOSIER, A. C., WELLS, G. F. & FRANCIS, C. A., 2012. Global Biodiversity of Aquatic Ammonia-Oxidizing Archaea is Partitioned by Habitat. *Front Microbiol*, **3**, 252.

BLACKWOOD, C., MARSH, T., KIM, S.-H. & PAUL, E., 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Appl Environ Microbiol*, **69**, 926–932.

BOLLMANN, A., BÄR-GILISSEN, M. J. & LAANBROEK, H. J., 2002. Growth at low ammonium concentrations and starvation response as potential factors involved in niche differentiation among ammonia-oxidizing bacteria. *Applied and environmental microbiology*, **68**(10), 4751–4757.

BORATYN, G. M., SCHÄFFER, A. A., AGARWALA, R., ALTSCHUL, S. F., LIPMAN, D. J. & MADDEN, T. L., 2012. Domain enhanced lookup time accelerated BLAST. *Biol Direct*, **7**, 12.

BORK, P., JENSEN, L. J., VON MERING, C., RAMANI, A. K., LEE, I. & MARCOTTE, E. M., 2004. Protein interaction networks from yeast to human. *Current opinion in structural biology*, **14**(3), 292–299.

BOROVOK, I., KREISBERG-ZAKARIN, R., YANKO, M., SCHREIBER, R., MYSLOVATI, M., ASLUND, F., HOLMGREN, A., COHEN, G. & AHARONOWITZ, Y., 2002. Streptomyces spp. contain class Ia and class II ribonucleotide reductases: expression analysis of the genes in vegetative growth. *Microbiology*, **148**(2), 391–404.

BOUSKILL, N. J., EVEILLARD, D., CHIEN, D., JAYAKUMAR, A. & WARD, B. B., 2011. Environmental factors determining ammonia-oxidizing organism distribution and diversity in marine environments. *Environ Microbiol*.

BRITTON, T., ANDERSON, C. L., JACQUET, D., LUNDQVIST, S. & BREMER, K., 2007. Estimating divergence times in large phylogenetic trees. *Systematic biology*, **56**(5), 741.

BROCHIER-ARMANET, C., BOUSSAU, B., GRIBALDO, S. & FORTERRE, P., 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*, **6**(3), 245–52.

BRYANT, J. A., LAMANNA, C., MORLON, H., KERKHOFF, A. J., ENQUIST, B. J. & GREEN, J. L., 2008. Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences USA*, **105**(Suppl. 1), 11505–11511.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, **10**(1), 421.

CAMP, A. H. & LOSICK, R., 2008. A novel pathway of intercellular signalling in Bacillus subtilis involves a protein with similarity to a component of type III secretion channels. *Molecular microbiology*, **69**(2), 402–417.

CAMP, A. H. & LOSICK, R., 2009. A feeding tube model for activation of a cell-specific transcription factor during sporulation in Bacillus subtilis. *Genes Dev*, **23**(8), 1014–24.

CANFIELD, D. E., ROSING, M. T. & BJERRUM, C., 2006. Early anaerobic metabolisms. *Philos Trans R Soc Lond B Biol Sci*, **361**(1474), 1819–34; discussion 1835–6.

CASTRESANA, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, **17**(4), 540–52.

CAVENDER-BARES, J., KOZAK, K. H., FINE, P. V. A. & KEMBEL, S. W., 2009. The merging of community ecology and phylogenetic biology. *Ecol Lett*, **12**(7), 693–715.

CHAI, Y., BEAUREGARD, P. B., VLAMAKIS, H., LOSICK, R. & KOLTER, R., 2012. Galactose Metabolism Plays a Crucial Role in Biofilm Formation by Bacillus subtilis. *MBio*, **3**(4).

CHAN, C. T. Y., PANG, Y. L. J., DENG, W., BABU, I. R., DYAVAIAH, M., BEGLEY, T. J. & DEDON, P. C., 2012. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat Commun*, **3**, 937.

CHATTOPADHYAY, M. K., PARK, M. H. & TABOR, H., 2008. Hypusine modification for growth is the major function of spermidine in Saccharomyces cerevisiae polyamine auxotrophs grown in limiting spermidine. *Proceedings of the National Academy of Sciences*, **105**(18), 6554.

CHEN, K. & PACHTER, L., 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS computational biology*, **1**(2), e24.

CHRISTOPHER, A. B., ARNDT, A., CUGINI, C. & DAVEY, M. E., 2010. A streptococcal effector protein that inhibits Porphyromonas gingivalis biofilm development. *Microbiology*, **156**(11), 3469–3477.

CHURCH, M. J., DELONG, E. F., DUCKLOW, H. W., KARNER, M. B., PRESTON, C. M. & KARL, D. M., 2003. Abundance and distribution of planktonic Archaea and Bacteria in the waters west of the Antarctic Peninsula. *Limnology and Oceanography*, 1893–1902.

CHURCH, M. J., WAI, B., KARL, D. M. & DELONG, E. F., 2010. Abundances of crenarchaeal amoA genes and transcripts in the Pacific Ocean. *Environ Microbiol*, **12**(3), 679–88.

CLAUSET, A., SHALIZI, C. R. & NEWMAN, M. E. J., 2007. Power-law distributions in empirical data. *Arxiv preprint arxiv:0706.1062*.

CODISPOTI, L. A., 2007. An oceanic fixed nitrogen sink exceeding 400 Tg N a^{-1} vs the concept of homeostasis in the fixed-nitrogen inventory. *Biogeosciences*, **4**(2), 233–253.

COLE, J. R., CHAI, B., FARRIS, R. J., WANG, Q., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., BANDELA, A. M., CARDENAS, E., GARRITY, G. M. & TIEDJE, J. M., 2007. The Ribosomal Database Project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*, **35**, D169–D172.

COMERON, J. M., 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol*, **41**(6), 1152–9.

CRAMER, P., BUSHNELL, D. A., FU, J., GNATT, A. L., MAIER-DAVIS, B., THOMPSON, N. E., BURGESS, R. R., EDWARDS, A. M., DAVID, P. R. & KORNBERG, R. D., 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, **288**(5466), 640–649.

CRAMER, P., BUSHNELL, D. A. & KORNBERG, R. D., 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, **292**(5523), 1863–76.

CROOKS, G. E., HON, G., CHANDONIA, J.-M. M. & BRENNER, S. E., 2004. WebLogo: a sequence logo generator. *Genome Res*, **14**(6), 1188–90.

DAMSTÉ, J. S. S., SCHOUTEN, S., HOPMANS, E. C., VAN DUIN, A. C. T. & GEENEVASEN, J. A. J., 2002. Crenarchaeol: the characteristic core glycerol dibiphytanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. *J Lipid Res*, **43**(10), 1641–51.

DE CORTE, D., YOKOKAWA, T., VARELA, M. M., AGOGUÉ, H. & HERNDL, G. J., 2008. Spatial distribution of Bacteria and Archaea and amoA gene copy numbers throughout the water column of the Eastern Mediterranean Sea. *ISME J*.

DE CORTE, D., YOKOKAWA, T., VARELA, M. M., AGOGUÉ, H. & HERNDL, G. J., 2009. Spatial distribution of Bacteria and Archaea and amoA gene copy numbers throughout the water column of the Eastern Mediterranean Sea. *ISME J*, **3**(2), 147–58.

DE LA TORRE, J. R., WALKER, C. B., INGALLS, A. E., KÖNNEKE, M. & STAHL, D. A., 2008. Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol*, **10**(3), 810–8.

DELONG, E. F., 2005. Microbial community genomics in the ocean. *Nature Reviews Microbiology*, **3**(6), 459–469.

DENG, H.-W. & FU, Y.-X., 1996. The Effects of Variable Mutation Rates Across Sites on the Phylogenetic Estimation of Effective Population Size or Mutation Rate of DNA Sequences. *Genetics*, **144**(3), 1271–1281.

DEUTSCH, C., SARMIENTO, J. L., SIGMAN, D. M., GRUBER, N. & DUNNE, J. P., 2007. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature*, **445**(7124), 163–7.

EDDY, S., 2010. HMMER3: a new generation of sequence homology search software. URL: http://hmmer. janelia. org. *Accessed*, **7**(25), 2010.

EDGAR, R., 2010a. Muscle: Protein multiple sequence alignment software.

EDGAR, R. C., 2010b. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), 2460–1.

EFRON, B., 2004. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, **99**(465), 96–104.

EFRON, B., 2005. *Local false discovery rates*. Division of Biostatistics, Stanford University.

EICHLER, J. & ADAMS, M. W. W., 2005. Posttranslational protein modification in Archaea. *Microbiology and Molecular Biology Reviews*, **69**(3), 393–425.

ERGUDER, T. H., BOON, N., WITTEBOLLE, L., MARZORATI, M. & VERSTRAETE, W., 2009. Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. *FEMS Microbiol Rev*, **33**(5), 855–69.

FAITH, D. P., 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**(1), 1–10.

FAUST, K. & RAES, J., 2012. Microbial interactions: from networks to models. *Nat Rev Microbiol*, **10**(8), 538–50.

FAY, J. C. & WU, C.-I., 2000. Hitchhiking Under Positive Darwinian Selection. *Genetics*, **155**(3), 1405–1413.

FELSENSTEIN, J., 1985. Phylogenies and the comparative method. *The American Naturalist*, **125**(1), 1–15.

FERNÀNDEZ-GUERRA, A., BUCHAN, A., MOU, X., CASAMAYOR, E. O. & GONZÁLEZ, J. M., 2010. T-RFPred: a nucleotide sequence size prediction tool for microbial community description based on terminal-restriction fragment length polymorphism chromatograms. *BMC Microbiol*, **10**, 262.

FERNÀNDEZ-GUERRA, A. & CASAMAYOR, E. O., 2012. Habitat-associated phylogenetic community patterns of microbial ammonia oxidizers. *PLoS One*, **7**(10), e47330.

FINN, R. D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J. E., GAVIN, O. L., GUNASEKARAN, P., CERIC, G. & FORSLUND, K., 2010. The Pfam protein families database. *Nucleic acids research*, **38**(suppl 1), D211–D222.

FITZJOHN, R. & DICKIE, I., 2007. TRAMPR: an R package for analysis and matching of terminal-restriction fragment length polymorphism (TRFLP) profiles. *Mol Ecol Notes*, **7**, 583–587.

FORTUNATO, S. & BARTHELEMY, M., 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**(1), 36.

FRANCIS, C. A., BEMAN, J. M. & KUYPERS, M. M. M., 2007. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J*, **1**(1), 19–27.

FRANCIS, C. A., ROBERTS, K. J., BEMAN, J. M., SANTORO, A. E. & OAKLEY, B. B., 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci U S A*, **102**(41), 14683–8.

FREESE, E. & YOSHIDA, A., 1965. The role of mutations in evolution. *Evolving genes and proteins. Academic Press, New York*, 341–355.

FU, Y. X., 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**(2), 915–25.

FU, Y. X. & LI, W. H., 1993a. Maximum likelihood estimation of population parameters. *Genetics*, **134**(4), 1261–70.

FU, Y. X. & LI, W. H., 1993b. Statistical tests of neutrality of mutations. *Genetics*, **133**(3), 693–709.

FUHRMAN, J. A., 2009. Microbial community structure and its functional implications. *Nature*, **459**(7244), 193–199.

FUHRMAN, J. A., MCCALLUM, K. & DAVIS, A. A., 1992. Novel major archaebacterial group from marine plankton. *Nature*, **356**(6365), 148–9.

FUHRMAN, J. A., SCHWALBACH, M. S. & STINGL, U., 2008. Proteorhodopsins: an array of physiological roles? *Nat Rev Microbiol*, **6**(6), 488–94.

FUKUDA, W., MORIMOTO, N., IMANAKA, T. & FUJIWARA, S., 2008. Agmatine is essential for the cell growth of Thermococcus kodakaraensis. *FEMS microbiology letters*, **287**(1), 113–120.

GALAND, P. E., GUTIÉRREZ-PROVECHO, C., MASSANA, R., GASOL, J. M. & CASAMAYOR, E. O., 2010. Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnology and Oceanography*, **55**(5), 2117–2125. ISSN 00243590.

GALLOWAY, J. N. & COWLING, E. B., 2002. Reactive nitrogen and the world: 200 years of change. *AMBIO: A Journal of the Human Environment*, **31**(2), 64–71.

GALPERIN, M. Y. & KOONIN, E. V., 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res*, **32**(18), 5452–63.

GEVERS, D., COHAN, F. M., LAWRENCE, J. G., SPRATT, B. G., COENYE, T., FEIL, E. J., STACKEBRANDT, E., VAN DE PEER, Y., VANDAMME, P. & THOMPSON, F. L., 2005. Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, **3**(9), 733–739.

GILBERT, J. A. & DUPONT, C. L., 2011. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci*, **3**, 347–71.

GILBERT, J. A., FIELD, D., HUANG, Y., EDWARDS, R., LI, W., GILNA, P. & JOINT, I., 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, **3**(8), e3042.

GOLDMAN, N. & YANG, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, **11**(5), 725–36.

GONZALEZ, J. M., SIMO, R., MASSANA, R., COVERT, J. S., CASAMAYOR, E. O., PEDROS-ALIO, C. & MORAN, M. A., 2000. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl Environ Microbiol*, **66**, 4237–4246.

GRAHAM, D. E., XU, H. & WHITE, R. H., 2002. Methanococcus jannaschii uses a pyruvoyl-dependent arginine decarboxylase in polyamine biosynthesis. *J Biol Chem*, **277**(26), 23500–7.

GRASSLY, N. C. & FRASER, C., 2008. Inserting proteins into the bacterial cytoplasmic membrane using the Sec and YidC translocases. *Nature Reviews Microbiology*, **6**(6), 477–87.

GRUBER, N. & GALLOWAY, J. N., 2008. An Earth-system perspective of the global nitrogen cycle. *Nature*, **451**(7176), 293–296.

GRÜNBERG, S., REICH, C., ZELLER, M. E., BARTLETT, M. S. & THOMM, M., 2010. Rearrangement of the RNA polymerase subunit H and the lower jaw in archaeal elongation complexes. *Nucleic Acids Res*, **38**(6), 1950–63.

GUBRY-RANGIN, C., HAI, B., QUINCE, C., ENGEL, M., THOMSON, B. C., JAMES, P., SCHLOTER, M., GRIFFITHS, R. I., PROSSER, J. I. & NICOL, G. W., 2011. Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc Natl Acad Sci U S A*, **108**(52), 21206–11.

HAGOPIAN, D. S. & RILEY, J. G., 1998. A closer look at the bacteriology of nitrification. *Aquacultural engineering*, **18**(4), 223–244.

HALLAM, S. J., MINCER, T. J., SCHLEPER, C., PRESTON, C. M., ROBERTS, K., RICHARDSON, P. M. & DELONG, E. F., 2006. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol*, **4**(4), e95.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, **5**(10), R245–9.

HARRINGTON, E. D., SINGH, A. H., DOERKS, T., LETUNIC, I., VON MERING, C., JENSEN, L. J., RAES, J. & BORK, P., 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proceedings of the National Academy of Sciences*, **104**(35), 13913.

HARRIS, M. A., CLARK, J., IRELAND, A., LOMAX, J., ASHBURNER, M., FOULGER, R., EILBECK, K., LEWIS, S., MARSHALL, B. & MUNGALL, C., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, **32**(Database issue), D258.

HATZENPICHLER, R., 2012. Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. *Appl Environ Microbiol*, **78**(21), 7501–10.

HAWKINS, T., CHITALE, M. & KIHARA, D., 2008. New paradigm in protein function prediction for large scale omics analysis. *Mol. BioSyst.*, **4**(3), 223–231.

HEINEMANN, I. U., SÖLL, D. & RANDAU, L., 2010. Transfer RNA processing in archaea: unusual pathways and enzymes. *FEBS Lett*, **584**(2), 303–9.

HELMUS, M. R., BLAND, T. J., WILLIAMS, C. K. & IVES, A. R., 2007. Phylogenetic Measures of Biodiversity. *Am Nat*, **169**(3).

HIRSCHMAN, L., CLARK, C., COHEN, K. B., MARDIS, S., LUCIANO, J., KOTTMANN, R., COLE, J., MARKOWITZ, V., KYRPIDES, N. & MORRISON, N., 2008. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS A Journal of Integrative Biology*, **12**(2), 129–136.

HOLMES, A. J., COSTELLO, A., LIDSTROM, M. E. & MURRELL, J. C., 1995. Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiol Lett*, **132**(3), 203–8.

HOLMES, D. E., MESTER, T., O'NEIL, R. A., PERPETUA, L. A., LARRAHONDO, M. J., GLAVEN, R., SHARMA, M. L., WARD, J. E., NEVIN, K. P. & LOVLEY, D. R., 2008. Genes for two multicopper proteins required for Fe(III) oxide reduction in Geobacter sulfurreducens have different expression patterns both in the subsurface and on energy-harvesting electrodes. *Microbiology*, **154**(Pt 5), 1422–35.

HOOPER, A. B. & TERRY, K. R., 1973. Specific Inhibitors of Ammonia Oxidation in Nitrosomonas. *Journal of Bacteriology*, **115**(2), 480–485.

HORRIGAN, S. G. & SPRINGER, A. L., 1990. Oceanic and Estuarine Ammonium Oxidation: Effects of Light. *Limnology and Oceanography*, **35**(2), pp. 479–482.

HU, A., JIAO, N. & ZHANG, C. L., 2011a. Community structure and function of planktonic Crenarchaeota: changes with depth in the South China Sea. *Microb Ecol*, **62**(3), 549–63.

HU, A., JIAO, N., ZHANG, R. & YANG, Z., 2011b. Niche partitioning of marine group I Crenarchaeota in the euphotic and upper mesopelagic zones of the East China Sea. *Appl Environ Microbiol*, **77**(21), 7469–78.

HUELSENBECK, J. P. & DYER, K. A., 2004. Bayesian estimation of positively selected sites. *J Mol Evol*, **58**(6), 661–72.

HUNTER, S., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A. ET AL., 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37**(Database issue), D211–5.

HURST, L. D., FEIL, E. J. & ROCHA, E. P. C., 2006. Protein evolution: causes of trends in amino-acid gain and loss. *Nature*, **442**(7105), E11–2; discussion E12.

IVES, A. & HELMUS, M., 2010. Phylogenetic metrics of community similarity. *The American Naturalist*, **176**(5), E128–E142.

JANSSON, B. P. M., MALANDRIN, L. & JOHANSSON, H. E., 2000. Cell cycle arrest in archaea by the hypusination inhibitor N 1-guanyl-1, 7-diaminoheptane. *Journal of bacteriology*, **182**(4), 1158–1161.

JAROSZEWSKI, L., LI, Z., KRISHNA, S. S., BAKOLITSA, C., WOOLEY, J., DEACON, A. M., WILSON, I. A. & GODZIK, A., 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol*, **7**(9), e1000205.

JIANG, X., LANGILLE, M. G. I., NECHES, R. Y., ELLIOT, M., LEVIN, S. A., EISEN, J. A., WEITZ, J. S. & DUSHOFF, J., 2012. Functional Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization. *PLoS One*, **7**(9), e43866.

JONES, C. M. & HALLIN, S., 2010. Ecological and evolutionary factors underlying global and local assembly of denitrifier communities. *ISME J*, **4**(5), 633–41.

JORDAN, G. & GOLDMAN, N., 2011. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*.

JOYCE, E. A., CHAN, K., SALAMA, N. R. & FALKOW, S., 2002. Redefining bacterial populations: a post-genomic reformation. *Nature Reviews: Genetics*, **3**(6), 462–473.

JUNIER, P., JUNIER, T. & WITZEL, K., 2008. TRiFLe, a program for in silico terminal restriction fragment length polymorphism analysis with user-defined sequence sets. *Appl Environ Microbiol*, **74**, 6452–6456.

KALANETRA, K. M., BANO, N. & HOLLIBAUGH, J. T., 2009. Ammonia-oxidizing Archaea in the Arctic Ocean and Antarctic coastal waters. *Environ Microbiol*, **11**(9), 2434–45.

KALIES, K. U., RAPOPORT, T. A. & HARTMANN, E., 1998. The β subunit of the Sec61 complex facilitates cotranslational protein transport and interacts with the signal peptidase during translocation. *The Journal of cell biology*, **141**(4), 887–894.

KANAGAWA, T., 2003. Bias and artifacts in multitemplate Polymerase Chain Reactions (PCR). *J Biosci Bioeng*, **96**, 317–323.

KANEHISA, M., ARAKI, M., GOTO, S., HATTORI, M., HIRAKAWA, M., ITOH, M., KATAYAMA, T., KAWASHIMA, S., OKUDA, S., TOKIMATSU, T. & YAMANISHI, Y., 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, **36**(suppl 1), D480–D484.

KANEKO, S., ROZENBLATT-ROSEN, O., MEYERSON, M. & MANLEY, J. L., 2007. The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3′ processing and transcription termination. *Genes Dev*, **21**(14), 1779–89.

KAPLAN, C. W. & KITTS, C. L., 2003. Variation between observed and true Terminal Restriction Fragment length is dependent on true TRF length and purine content. *J Microbiol Methods*, **54**, 121–125.

KARL, D. M., 2007. Microbial oceanography: paradigms, processes and promise. *Nat Rev Microbiol*, **5**(10), 759–69.

KARNER, M. B., DELONG, E. F. & KARL, D. M., 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature*, **409**(6819), 507–10.

KATOH, K., KUMA, K.-I., MIYATA, T. & TOH, H., 2005. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform*, **16**(1), 22–33.

KEMBEL, S. W., COWAN, P. D., HELMUS, M. R., CORNWELL, W. K., MORLON, H., ACKERLY, D. D., BLOMBERG, S. P. & WEBB, C. O., 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**(11), 1463.

KENT, A. D., SMITH, D. J., BENSON, B. J. & TRIPLETT, E. W., 2003. Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities. *Appl Environ Microbiol*, **69**, 6768–6776.

KHANIN, R. & WIT, E., 2006. How scale-free are biological networks. *Journal of computational biology*, **13**(3), 810–818.

KIM, B. K., JUNG, M.-Y. Y., YU, D. S., PARK, S.-J. J., OH, T. K., RHEE, S.-K. K. & KIM, J. F., 2011. Genome sequence of an ammonia-oxidizing soil archaeon, "Candidatus Nitrosoarchaeum koreensis" MY1. *J Bacteriol*, **193**(19), 5539–40.

KIM, K. K., HUNG, L. W., YOKOTA, H., KIM, R. & KIM, S. H., 1998. Crystal structures of eukaryotic translation initiation factor 5A from Methanococcus jannaschii at 1.8 Å resolution. *Proceedings of the National Academy of Sciences*, **95**(18), 10419.

KIMURA, M., 1968. Evolutionary Rate at the Molecular Level. *Nature*, **217**(5129), 624–626.

KIMURA, M., ISHIGURO, A. & ISHIHAMA, A., 1997. RNA polymerase II subunits 2, 3, and 11 form a core subassembly with DNA binding activity. *Journal of Biological Chemistry*, **272**(41), 25851–25855.

KINCH, L. N. & SAIER JR, M. H., 2002. Sec61βa component of the archaeal protein secretory system. *Trends Biochem Sci*, **27**, 170–171.

KISHINO, H. & HASEGAWA, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*, **29**(2), 170–9.

KÖNNEKE, M., BERNHARD, A. E., DE LA TORRE, J. R., WALKER, C. B., WATERBURY, J. B. & STAHL, D. A., 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**(7058), 543–6.

KOONIN, E. V., 2007. Metagenomic sorcery and the expanding protein universe. *Nature Biotechnology*, **25**(5), 540–542.

KREUZER, K. N., 2005. Interplay Between DNA Replication and Recombination in Prokaryotes. *Annual Review of Microbiology*, **59**(1), 43–67. PMID: 15792496.

KRYAZHIMSKIY, S. & PLOTKIN, J. B., 2008. The population genetics of dN/dS. *PLoS Genet*, **4**(12), e1000304.

KRZYWINSKI, M., BIROL, I., JONES, S. J. & MARRA, M. A., 2012. Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics*, **13**(5), 627–644.

KYRPIDES, N. C. & WOESE, C. R., 1998. Universally conserved translation initiation factors. *Proceedings of the National Academy of Sciences*, **95**(1), 224.

LAM, P., JENSEN, M. M., LAVIK, G., MCGINNIS, D. F., MÜLLER, B., SCHUBERT, C. J., AMANN, R., THAMDRUP, B. & KUYPERS, M. M. M., 2007. Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea. *Proc Natl Acad Sci U S A*, **104**(17), 7104–9.

LAM, P., LAVIK, G., JENSEN, M. M., VAN DE VOSSENBERG, J., SCHMID, M., WOEBKEN, D., GUTIÉRREZ, D., AMANN, R., JETTEN, M. S. M. & KUYPERS, M. M. M., 2009. Revising the nitrogen cycle in the Peruvian oxygen minimum zone. *Proceedings of the National Academy of Sciences*.

LEE, S. H., KIM, P.-J. J., AHN, Y.-Y. Y. & JEONG, H., 2010. Googling social interactions: web search engine based social network construction. *PLoS One*, **5**(7), e11233.

LEHTOVIRTA, L. E., PROSSER, J. I. & NICOL, G. W., 2009. Soil pH regulates the abundance and diversity of Group 1.1c Crenarchaeota. *FEMS Microbiol Ecol*, **70**(3), 367–76.

LEHTOVIRTA-MORLEY, L. E., STOECKER, K., VILCINSKAS, A., PROSSER, J. I. & NICOL, G. W., 2011. Cultivation of an obligate acidophilic ammonia oxidizer from a nitrifying acid soil. *Proc Natl Acad Sci U S A*, **108**(38), 15892–7.

LEININGER, S., URICH, T., SCHLOTER, M., SCHWARK, L., QI, J., NICOL, G. W., PROSSER, J. I., SCHUSTER, S. C. & SCHLEPER, C., 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**(7104), 806–9.

LEIPE, D. D., ARAVIND, L., GRISHIN, N. V. & KOONIN, E. V., 2000. The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome research*, **10**(1), 5–16.

LETUNIC, I. & BORK, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**(1), 127–8.

LEVIN, S. A., 2003. Complex adaptive systems: exploring the known, the unknown and the unknowable. *Bulletin of the American Mathematical Society*, **40**(1), 3–20.

LI, W. & GODZIK, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–9.

LI, W., WOOLEY, J. C. & GODZIK, A., 2008a. Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE*, **3**(10), e3375.

LI, W. H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*, **36**(1), 96–9.

LI, W. H., WU, C. I. & LUO, C. C., 1985a. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, **2**(2), 150–174.

LI, W. H., WU, C. I. & LUO, C. C., 1985b. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, **2**(2), 150–74.

LI, Y. F., COSTELLO, J. C., HOLLOWAY, A. K. & HAHN, M. W., 2008b. "REVERSE ECOLOGY" AND THE POWER OF POPULATION GENOMICS. *Evolution*, **62**(12), 2984–2994.

LITTLE, E., BORK, P. & DOOLITTLE, R. F., 1994. Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *Journal of molecular evolution*, **39**(6), 631–643.

LIU, W.-T. . T., MARSH, T. L., CHENG, H. & FORNEY, L. J., 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol*, **63**, 4516–4522.

LOZUPONE, C. & KNIGHT, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, **71**(12), 8228.

MA, G. & RD, J., 1996a. Photoinhibition of marine nitrifying bacteria. I. Wavelength-dependent response. *Marine Ecology Progress Series*, **141**, 183–192.

MA, G. & RD, J., 1996b. Photoinhibition of marine nitrifying bacteria. II. Dark recovery after monochromatic or polychromatic irradiation. *Marine Ecology Progress Series*, **141**, 193–198.

MA, L., LU, H., SPRINKLE, A., PARSEK, M. R. & WOZNIAK, D. J., 2007. Pseudomonas aeruginosa Psl is a galactose-and mannose-rich exopolysaccharide. *Journal of bacteriology*, **189**(22), 8353–8356.

MANDAL, D., KÖHRER, C., SU, D., RUSSELL, S. P., KRIVOS, K., CASTLEBERRY, C. M., BLUM, P., LIMBACH, P. A., SÖLL, D. & RAJBHANDARY, U. L., 2010. Agmatidine, a modified cytidine in the anticodon of archaeal tRNA(Ile), base pairs with adenosine but not with guanosine. *Proc Natl Acad Sci U S A*, **107**(7), 2872–7.

MARGOLIS, J. J., EL-ETR, S., JOUBERT, L. M., MOORE, E., ROBISON, R., RASLEY, A., SPORMANN, A. M. & MONACK, D. M., 2010. Contributions of Francisella tularensis subsp. novicida chitinases and Sec secretion system to biofilm formation on chitin. *Applied and environmental microbiology*, **76**(2), 596–608.

MARSH, T., 1999. Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Curr Opin Microbiol*, **2**, 323–327.

MARSH, T., 2005. Culture-independent microbial community analysis with terminal restriction fragment length polymorphism. *Methods Enzymol*, **397**, 308–329.

MARSH, T., SAXMAN, P., COLE, J. & TIEDJE, J., 2000. Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. *Appl Environ Microbiol*, **66**, 3616–3620.

MARTENS-HABBENA, W., BERUBE, P. M., URAKAWA, H., DE LA TORRE, J. R. & STAHL, D. A., 2009. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature*, **461**(7266), 976–979.

MARTENS-HABBENA, W. & STAHL, D. A., 2011. Nitrogen metabolism and kinetics of ammonia-oxidizing archaea. *Methods Enzymol*, **496**, 465–87.

MARTINY, J. B. H., BOHANNAN, B. J. M., BROWN, J. H., COLWELL, R. K., FUHRMAN, J. A., GREEN, J. L., HORNER-DEVINE, M. C., KANE, M., KRUMINS, J. A. & KUSKE, C. R., 2006. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**(2), 102–112.

MASHHADI, Z., ZHANG, H., XU, H. & WHITE, R. H., 2008. Identification and characterization of an archaeon-specific riboflavin kinase. *Journal of bacteriology*, **190**(7), 2615–2618.

MCTAVISH, H., FUCHS, J. A. & HOOPER, A. B., 1993. Sequence of the gene coding for ammonia monooxygenase in Nitrosomonas europaea. *J Bacteriol*, **175**(8), 2436–44.

MEHTA, T., CHILDERS, S. E., GLAVEN, R., LOVLEY, D. R. & MESTER, T., 2006. A putative multicopper protein secreted by an atypical type II secretion system involved in the reduction of insoluble electron acceptors in Geobacter sulfurreducens. *Microbiology*, **152**(Pt 8), 2257–64.

MENDUM, T. A., SOCKETT, R. E. & HIRSCH, P. R., 1999. Use of molecular and isotopic techniques to monitor the response of autotrophic ammonia-oxidizing populations of the beta subdivision of the class proteobacteria in arable soils to nitrogen fertilizer. *Appl Environ Microbiol*, **65**(9), 4155–62.

MERBT, S. N., STAHL, D. A., CASAMAYOR, E. O., MARTÍ, E., NICOL, G. W. & PROSSER, J. I., 2012. Differential photoinhibition of bacterial and archaeal ammonia oxidation. *FEMS Microbiol Lett*, **327**(1), 41–6.

MES, T. H. M., 2008. Microbial diversity–insights from population genetics. *Environ Microbiol*, **10**(1), 251–64.

MINAKHIN, L., BHAGAT, S., BRUNNING, A., CAMPBELL, E. A., DARST, S. A., EBRIGHT, R. H. & SEVERINOV, K., 2001. Bacterial RNA polymerase subunit ω and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proceedings of the National Academy of Sciences*, **98**(3), 892.

MINCER, T. J., CHURCH, M. J., TAYLOR, L. T., PRESTON, C., KARL, D. M. & DELONG, E. F., 2007. Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environ Microbiol*, **9**(5), 1162–75.

MISAWA, K. & TAJIMA, F., 1997. Estimation of the Amount of DNA Polymorphism When the Neutral Mutation Rate Varies Among Sites. *Genetics*, **147**(4), 1959–1964.

MIYAO, T. & WOYCHIK, N. A., 1998. RNA polymerase subunit RPB5 plays a role in transcriptional activation. *Proceedings of the National Academy of Sciences*, **95**(26), 15281.

MIYATA, T. & YASUNAGA, T., 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, **16**(1), 23–36.

MOLINA, V., ULLOA, O., FARÍAS, L., URRUTIA, H., RAMÍREZ, S., JUNIER, P. & WITZEL, K.-P. P., 2007. Ammonia-oxidizing beta-proteobacteria from the oxygen minimum zone off northern Chile. *Appl Environ Microbiol*, **73**(11), 3547–55.

MOOERS, A. & HEARD, S., 1997. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, **72**(1), 31–54.

MOORE, M. J. & PROUDFOOT, N. J., 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, **136**(4), 688–700.

MORI, T., SAVELIEV, S. V., XU, Y., STAFFORD, W. F., COX, M. M., INMAN, R. B. & JOHNSON, C. H., 2002. Circadian clock protein KaiC forms ATP-dependent hexameric rings and binds DNA. *Proceedings of the National Academy of Sciences*, **99**(26), 17203.

MOSIER, A. C. & FRANCIS, C. A., 2011. Determining the distribution of marine and coastal ammonia-oxidizing archaea and bacteria using a quantitative approach. *Methods Enzymol*, **486**, 205–21.

MOU, X., MORAN, M. A., STEPANAUSKAS, R., GONZALEZ, J. M. & HODSON, R. E., 2005. Flow-cytometric cell sorting and subsequent molecular analyses for culture-independent identification of bacterioplankton involved in dimethylsulfoniopropionate transformations. *Appl Environ Microbiol*, **71**, 1405–1416.

MURRELL, B., WERTHEIM, J. O., MOOLA, S., WEIGHILL, T., SCHEFFLER, K. & KOSAKOVSKY POND, S. L., 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*, **8**(7), e1002764.

MUSE, S. V. & GAUT, B. S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, **11**(5), 715–24.

MUSSMANN, M., BRITO, I., PITCHER, A., SINNINGHE DAMSTÉ, J. S., HATZENPICHLER, R., RICHTER, A., NIELSEN, J. L., NIELSEN, P. H., MÜLLER, A., DAIMS, H., WAGNER, M. & HEAD, I. M., 2011. Thaumarchaeotes abundant in refinery nitrifying sludges express amoA but are not obligate autotrophic ammonia oxidizers. *Proc Natl Acad Sci U S A*, **108**(40), 16771–6.

NEI, M., 2005. Selectionism and neutralism in molecular evolution. *Mol Biol Evol*, **22**(12), 2318–42.

NEI, M. & GOJOBORI, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, **3**(5), 418–26.

NEI, M. & GRAUR, D., 1984. Extent of Protein Polymorphism and the Neutral Mutation Theory. *Evolutionary Biology*, **17**, 73–118.

NICOL, G. W., LEININGER, S., SCHLEPER, C. & PROSSER, J. I., 2008. The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. *Environ Microbiol*, **10**(11), 2966–78.

NICOL, G. W. & SCHLEPER, C., 2006. Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? *Trends Microbiol*, **14**(5), 207–12.

NIELSEN, R. & YANG, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**(3), 929–936.

NISHIMASU, H., FUSHINOBU, S., SHOUN, H. & WAKAGI, T., 2004. The first crystal structure of the novel class of fructose-1,6-bisphosphatase present in thermophilic archaea. *Structure*, **12**(6), 949–59.

NOLD, S. C., ZHOU, J., DEVOL, A. H. & TIEDJE, J. M., 2000. Pacific Northwest marine sediments contain ammonia-oxidizing bacteria in the beta subdivision of the Proteobacteria. *Appl Environ Microbiol*, **66**(10), 4532–5.

OCHSENREITER, T., SELEZI, D., QUAISER, A., BONCH-OSMOLOVSKAYA, L. & SCHLEPER, C., 2003. Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environmental Microbiology*, **5**(9), 787–797.

OHTA, T., 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.*, **23**, 263–286.

OLSON, R. J., 1981. 15N tracer studies of the primary nitrite maximum. *J. Mar. Res.*, **39**, 203–226.

OSAWA, T., KIMURA, S., TERASAKA, N., INANAGA, H., SUZUKI, T. & NUMATA, T., 2011. Structural basis of tRNA agmatinylation essential for AUA codon decoding. *Nat Struct Mol Biol*, **18**(11), 1275–80.

PAMILO, P. & BIANCHI, N. O., 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol Evol*, **10**(2), 271–81.

PARADIS, E., CLAUDE, J. & STRIMMER, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2), 289–90.

PARK, E. & RAPOPORT, T. A., 2011. Mechanisms of Sec61/SecY-Mediated Protein Translocation Across Membranes. *Annual review of biophysics*.

PEAT, T. S., NEWMAN, J., WALDO, G. S., BERENDZEN, J. & TERWILLIGER, T. C., 1998. Structure of translation initiation factor 5A from Pyrobaculum aerophilum at 1.75 Å resolution. *Structure*, **6**(9), 1207–1214.

PENN, O., PRIVMAN, E., LANDAN, G., GRAUR, D. & PUPKO, T., 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular biology and evolution*, **27**(8), 1759–1767.

PESTER, M., RATTEI, T., FLECHL, S., GRÖNGRÖFT, A., RICHTER, A., OVERMANN, J., REINHOLD-HUREK, B., LOY, A. & WAGNER, M., 2012. amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. *Environ Microbiol*, **14**(2), 525–39.

PESTER, M., SCHLEPER, C. & WAGNER, M., 2011. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol*, **14**(3), 300–6.

PINHASSI, J., SIMO, R., GONZALEZ, J., VILA, M., ALONSO-SAEZ, L., KIENE, R., MORAN, M. & PEDROS-ALIO, C., 2005. Dimethylsulfoniopropionate turnover is linked to the composition and dynamics of the bacterioplankton assemblage during a microcosm phytoplankton bloom. *Appl Environ Microbiol*, **71**, 7650–7660.

PITCHER, A., VILLANUEVA, L., HOPMANS, E. C., SCHOUTEN, S., REICHART, G.-J. J. & SINNINGHE DAMSTÉ, J. S., 2011. Niche segregation of ammonia-oxidizing archaea and anammox bacteria in the Arabian Sea oxygen minimum zone. *ISME J*, **5**(12), 1896–904.

POND, S. L. K. & FROST, S. D. W., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, **22**(5), 1208–1222.

POND, S. L. K., FROST, S. D. W., GROSSMAN, Z., GRAVENOR, M. B., RICHMAN, D. D. & BROWN, A. J. L., 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS computational biology*, **2**(6), e62.

POND, S. L. K., FROST, S. D. W. & MUSE, S. V., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**(5), 676–9.

POND, S. L. K., MURRELL, B., FOURMENT, M., FROST, S. D. W., DELPORT, W. & SCHEFFLER, K., 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution*, **28**(11), 3033–3043.

PRIVMAN, E., PENN, O. & PUPKO, T., 2011. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Molecular Biology and Evolution*.

PROSSER, J. I. & NICOL, G. W., 2008. Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment. *Environmental Microbiology*, **10**(11), 2931–2941.

PRUESSE, E., QUAST, C., KNITTEL, K., FUCHS, B. M., LUDWIG, W., PEPLIES, J. & GLÖCKNER, F. O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, **35**(21), 7188–7196.

PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., BATEMAN, A. & FINN, R. D., 2012. The Pfam protein families database. *Nucleic Acids Res*, **40**(Database issue), D290–301.

PYBUS, O. G. & HARVEY, P. H., 2000. Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **267**(1459), 2267.

REIGSTAD, L. J., RICHTER, A., DAIMS, H., URICH, T., SCHWARK, L. & SCHLEPER, C., 2008. Nitrification in terrestrial hot springs of Iceland and Kamchatka. *FEMS Microbiol Ecol*, **64**(2), 167–74.

REVA, O. & TÜMMLER, B., 2008. Think big–giant genes in bacteria. *Environ Microbiol*, **10**(3), 768–77.

REZNICK, D. N. & RICKLEFS, R. E., 2009. Darwin's bridge between microevolution and macroevolution. *Nature*, **457**(7231), 837–42.

RICE, P., LONGDEN, I. & BLEASBY, A., 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet*, **16**, 276–277.

RICKE, P., KOLB, S. & BRAKER, G., 2005. Application of a newly developed ARB software-integrated tool for in silico terminal restriction fragment length polymorphism analysis reveals the dominance of a novel pmoA cluster in a forest soil. *Appl Environ Microbiol*, **71**, 1671–1673.

ROBERTS, R. J., VINCZE, T., POSFAI, J. & MACELIS, D., 2005. REBASE–restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*, **33**, D230–D232.

ROCHA, E. P. C., SMITH, J. M., HURST, L. D., HOLDEN, M. T. G., COOPER, J. E., SMITH, N. H. & FEIL, E. J., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*, **239**(2), 226–35.

ROCKSTRÖM, J., STEFFEN, W., NOONE, K., PERSSON, Å., CHAPIN, F. S., LAMBIN, E. F., LENTON, T. M., SCHEFFER, M., FOLKE, C. & SCHELLNHUBER, H. J., 2009. A safe operating space for humanity. *Nature*, **461**(7263), 472–475.

ROGERS, A. D., 2000. The role of the oceanic oxygen minima in generating biodiversity in the deep sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, **47**(1), 119–148.

ROSCH, C. & BOTHE, H., 2005. Improved assessment of denitrifying, N2-fixing, and total-community bacteria by terminal restriction fragment length polymorphism analysis using multiple restriction enzymes. *Appl Environ Microbiol*, **71**, 2026–2035.

ROTTHAUWE, J. H., WITZEL, K. P. & LIESACK, W., 1997. The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl Environ Microbiol*, **63**(12), 4704–12.

RUSCH, D. B., HALPERN, A. L., SUTTON, G., HEIDELBERG, K. B., WILLIAMSON, S., YOOSEPH, S., WU, D., EISEN, J. A., HOFFMAN, J. M. & REMINGTON, K., 2007. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, **5**(3), e77.

RYDER, C., BYRD, M. & WOZNIAK, D. J., 2007. Role of polysaccharides in Pseudomonas aeruginosa biofilm development. *Current opinion in microbiology*, **10**(6), 644–648.

SANTORO, A. E., CASCIOTTI, K. L. & FRANCIS, C. A., 2010. Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. *Environ Microbiol*, **12**(7), 1989–2006.

SAUDER, L. A., ENGEL, K., STEARNS, J. C., MASELLA, A. P., PAWLISZYN, R. & NEUFELD, J. D., 2011. Aquarium nitrification revisited: thaumarchaeota are the dominant ammonia oxidizers in freshwater aquarium biofilters. *PLoS One*, **6**(8), e23281.

SAUDER, L. A., PETERSE, F., SCHOUTEN, S. & NEUFELD, J. D., 2012. Low-ammonia niche of ammonia-oxidizing archaea in rotating biological contactors of a municipal wastewater treatment plant. *Environ Microbiol*.

SCHÄFER, G., ENGELHARD, M. & MÜLLER, V., 1999. Bioenergetics of the Archaea. *Microbiology and molecular biology reviews*, **63**(3), 570–620.

SCHÄFER, J., OPGEN-RHEIN, R. & STRIMMER, K., 2001. Reverse engineering genetic networks using the GeneNet package. *Journal of the American Statistical Association*, **96**, 1151–1160.

SCHÄFER, J. & STRIMMER, K., 2005a. Learning Large-Scale Graphical Gaussian Models from Genomic Data. In *AIP Conference Proceedings*, vol. 776, 263.

SCHÄFER, J. & STRIMMER, K., 2005b. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, **4**(1), 32.

SCHEFFLER, K., MARTIN, D. P. & SEOIGHE, C., 2006. Robust inference of positive selection from recombining coding sequences. *Bioinformatics*, **22**(20), 2493–2499.

SCHIMEL, D. S., HOUSE, J. I., HIBBARD, K. A., BOUSQUET, P., CIAIS, P. ET AL., 2001. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, **414**(6860), 169–72.

SCHOLZ, M. B., LO, C.-C. C. & CHAIN, P. S. G., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*, **23**(1), 9–15.

SEITZINGER, S., HARRISON, J. A., BÖHLKE, J. K., BOUWMAN, A. F., LOWRANCE, R., PETERSON, B., TOBIAS, C. & VAN DRECHT, G., 2006. Denitrification across landscapes and waterscapes: a synthesis. *Ecol Appl*, **16**(6), 2064–90.

SESHADRI, R., KRAVITZ, S. A., SMARR, L., GILNA, P. & FRAZIER, M., 2007. CAMERA: a community resource for metagenomics. *PLoS Biol*, **5**(3), e75.

SHIEMKE, A. K. . K., ARP, D. J. . J. & SAYAVEDRA-SOTO, L. A. . A., 2004. Inhibition of Membrane-Bound Methane Monooxygenase and Ammonia Monooxygenase by Diphenyliodonium: Implications for Electron Transfer. *Journal of Bacteriology*, **186**(4), 928–937.

SHIMODAIRA, H., 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, **51**(3), 492–508.

SHIMODAIRA, H. & HASEGAWA, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, **16**, 1114–1116.

SHIMODAIRA, H. & HASEGAWA, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**(12), 1246–1247.

SHRINER, D., NICKLE, D. C., JENSEN, M. A. & MULLINS, J. I., 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res*, **81**(2), 115–21.

SHYU, C., SOULE, T., BENT, S. J., FOSTER, J. A. & FORNEY, L. J., 2007. MiCA: a web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphisms of 16S and 18S rRNA genes. *Microb Ecol*, **53**, 562–570.

SINIGALLIANO, C. D., KUHN, D. N. & JONES, R. D., 1995. Amplification of the amoA gene from diverse species of ammonium-oxidizing bacteria and from an indigenous bacterial population from seawater. *Appl Environ Microbiol*, **61**(7), 2702–6.

SINTES, E., BERGAUER, K., DE CORTE, D., YOKOKAWA, T. & HERNDL, G. J., 2012. Archaeal amoA gene diversity points to distinct biogeography of ammonia-oxidizing Crenarchaeota in the ocean. *Environ Microbiol*.

SMIL, V., 2004. *Enriching the earth: Fritz Haber, Carl Bosch, and the transformation of world food production*. MIT press.

SMITH, T. F. & WATERMAN, M. S., 1981. Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–197.

SPANG, A., HATZENPICHLER, R., BROCHIER-ARMANET, C., RATTEI, T., TISCHLER, P., SPIECK, E., STREIT, W., STAHL, D. A., WAGNER, M. & SCHLEPER, C., 2010. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol*, **18**(8), 331–40.

STAHL, D. A. & DE LA TORRE, J. R., 2012. Physiology and diversity of ammonia-oxidizing archaea. *Annu Rev Microbiol*, **66**, 83–101.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A. ET AL., 2002. The bioperl toolkit: Perl modules for the life sciences. *Genome Res*, **12**, 1611–1618.

STAMATAKIS, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–90.

STAMATAKIS, A., LUDWIG, T. & MEIER, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**(4), 456–63.

STEPHEN, J. R., CHANG, Y. J., MACNAUGHTON, S. J., KOWALCHUK, G. A., LEUNG, K. T., FLEMMING, C. A. & WHITE, D. C., 1999. Effect of toxic metals on indigenous soil beta-subgroup proteobacterium ammonia oxidizer community structure and protection against toxicity by inoculated metal-resistant bacteria. *Appl Environ Microbiol*, **65**(1), 95–101.

STERNER, R. W. & ELSER, J. J., 2002. *Ecological stoichiometry: the biology of elements from molecules to the biosphere*. Princeton University Press.

SUZUKI, Y. & GOJOBORI, T., 1999. A method for detecting positive selection at single amino acid sites. *Molecular Biology And Evolution*, **16**(10), 1315–1328.

SUZUKI, Y. & NEI, M., 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. *Molecular biology and evolution*, **21**(5), 914–921.

TAJIMA, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**(3), 585–95.

TAJIMA, F., 1996. The Amount of DNA Polymorphism Maintained in a Finite Population When the Neutral Mutation Rate Varies Among Sites. *Genetics*, **143**(3), 1457–1465.

TEAM, R., 2010. R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria*, (01/19).

TOURNA, M., STIEGLMEIER, M., SPANG, A., KÖNNEKE, M., SCHINTLMEISTER, A., URICH, T., ENGEL, M., SCHLOTER, M., WAGNER, M., RICHTER, A. & SCHLEPER, C., 2011. Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci U S A*, **108**(20), 8420–5.

TREUSCH, A. H., LEININGER, S., KLETZIN, A., SCHUSTER, S. C., KLENK, H.-P. P. & SCHLEPER, C., 2005. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ Microbiol*, **7**(12), 1985–95.

ULLOA, O., CANFIELD, D. E., DELONG, E. F., LETELIER, R. M. & STEWART, F. J., 2012. Microbial oceanography of anoxic oxygen minimum zones. *Proc Natl Acad Sci U S A*, **109**(40), 15996–6003.

VALENTINE, D. L., 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology*, **5**(4), 316–323.

VAN LOOSDRECHT, M. C. M. & JETTEN, M. S. M., 1998. Microbiological conversions in nitrogen removal. *Water Science and Technology*, **38**(1), 1–7.

VARELA, M. M., VAN AKEN, H. M., SINTES, E. & HERNDL, G. J., 2008. Latitudinal trends of Crenarchaeota and Bacteria in the meso- and bathypelagic water masses of the Eastern North Atlantic. *Environ Microbiol*, **10**(1), 110–24.

VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D. ET AL., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**(5667), 66–74.

VERHAMME, D. T., PROSSER, J. I. & NICOL, G. W., 2011. Ammonia concentration determines differential growth of ammonia-oxidising archaea and bacteria in soil microcosms. *The ISME journal*, **5**(6), 1067–1071.

VITOUSEK, P. M., HÄTTENSCHWILER, S., OLANDER, L. & ALLISON, S., 2002. Nitrogen and nature. *AMBIO: A Journal of the Human Environment*, **31**(2), 97–101.

WAGNER, A., 2008. Nat Rev Genet.

WAGNER, S. & KLUG, G., 2007. An archaeal protein with homology to the eukaryotic translation initiation factor 5A shows ribonucleolytic activity. *J Biol Chem*, **282**(19), 13966–76.

WALKER, C. B., DE LA TORRE, J. R., KLOTZ, M. G., URAKAWA, H., PINEL, N. ET AL., 2010. Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci U S A*, **107**(19), 8818–23.

WEBB, C., 2000. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, **156**(2), 145–155.

WEBB, C., ACKERLY, D., MCPEEK, M. & DONOGHUE, M., 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505.

WERNER, F. & GROHMANN, D., 2011. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol*, **9**(2), 85–98.

WESTBLADE, L. F., CAMPBELL, E. A., PUKHRAMBAM, C., PADOVAN, J. C., NICKELS, B. E., LAMOUR, V. & DARST, S. A., 2010. Structural basis for the bacterial transcription-repair coupling factor/RNA polymerase interaction. *Nucleic Acids Res*, **38**(22), 8357–69.

WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K. ET AL., 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **36**(Database issue), D13–21.

WHITAKER, R. J. & BANFIELD, J. F., 2006. Population genomics in natural microbial communities. *Trends Ecol Evol*, **21**(9), 508–16.

WHITTAKER, J., 1990. *Graphical models in applied multivariate statistics*, vol. 16. Wiley New York.

WOLFF, E. C., KANG, K. R., KIM, Y. S. & PARK, M. H., 2007. Posttranslational synthesis of hypusine: evolutionary progression and specificity of the hypusine modification. *Amino Acids*, **33**(2), 341–350.

WUCHTER, C., ABBAS, B., COOLEN, M. J. L., HERFORT, L., VAN BLEIJSWIJK, J., TIMMERS, P., STROUS, M., TEIRA, E., HERNDL, G. J., MIDDELBURG, J. J., SCHOUTEN, S. & SINNINGHE DAMSTÉ, J. S., 2006. Archaeal nitrification in the ocean. *Proc Natl Acad Sci U S A*, **103**(33), 12317–22.

WUCHTY, S., 2001. Scale-free behavior in protein domain networks. *Molecular biology and evolution*, **18**(9), 1694–1702.

YAKIMOV, M. M., CONO, V. L. & DENARO, R., 2009. A first insight into the occurrence and expression of functional amoA and accA genes of autotrophic and ammonia-oxidizing bathypelagic Crenarchaeota of Tyrrhenian Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, **56**(11-12), 748–754.

YAKIMOV, M. M., LA CONO, V., DENARO, R., D'AURIA, G., DECEMBRINI, F., TIMMIS, K. N., GOLYSHIN, P. N. & GIULIANO, L., 2007. Primary producing prokaryotic communities of brine, interface and seawater above the halocline of deep anoxic lake L'Atalante, Eastern Mediterranean Sea. *ISME J*, **1**(8), 743–55.

YAN, J., HAAIJER, S. C. M., OP DEN CAMP, H. J. M., VAN NIFTRIK, L., STAHL, D. A., KÖNNEKE, M., RUSH, D., SINNINGHE DAMSTÉ, J. S., HU, Y. Y. & JETTEN, M. S. M., 2012. Mimicking the oxygen minimum zones: stimulating interaction of aerobic archaeal and anaerobic bacterial ammonia oxidizers in a laboratory-scale model system. *Environ Microbiol*.

YANG, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8), 1586–91.

YANG, Z. & NIELSEN, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**(1), 32–43.

YANG, Z. & NIELSEN, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**(6), 908–917.

YANG, Z., NIELSEN, R., GOLDMAN, N. & PEDERSEN, A. M. K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1), 431–449.

YANG, Z. & SWANSON, W. J., 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*, **19**(1), 49–57.

YAYANOS, A. A., 1995. Microbiology to 10,500 meters in the deep sea. *Annual Reviews in Microbiology*, **49**(1), 777–805.

YE, W., LIU, X., LIN, S., TAN, J., PAN, J., LI, D. & YANG, H., 2009. The vertical distribution of bacterial and archaeal communities in the water and sediment of Lake Taihu. *FEMS Microbiology Ecology*, **70**(2), 263–276.

YEE, A., BOOTH, V., DHARAMSI, A., ENGEL, A., EDWARDS, A. M. & ARROWSMITH, C. H., 2000. Solution structure of the RNA polymerase subunit RPB5 from Methanobacterium thermoautotrophicum. *Proceedings of the National Academy of Sciences*, **97**(12), 6311.

YILMAZ, P., IVERSEN, M. H., HANKELN, W., KOTTMANN, R., QUAST, C. & GLÖCKNER, F. O., 2012. Ecological structuring of bacterial and archaeal taxa in surface ocean waters. *FEMS Microbiol Ecol*, **81**(2), 373–85.

YOOL, A., MARTIN, A. P., FERNÁNDEZ, C. & CLARK, D. R., 2007. The significance of nitrification for oceanic new production. *Nature*, **447**(7147), 999–1002.

YOOSEPH, S., LI, W. & SUTTON, G., 2008. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC bioinformatics*, **9**(1), 182.

ZANELLI, C. F. & VALENTINI, S. R., 2007. Is there a role for eIF5A in translation? *Amino Acids*, **33**(2), 351–8.

ZÁRATE, S., POND, S. L. K., SHAPSHAK, P. & FROST, S. D. W., 2007. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. *J Virol*, **81**(12), 6643–51.

ZHANG, C. L., YE, Q., HUANG, Z., LI, W., CHEN, J., SONG, Z., ZHAO, W., BAGWELL, C., INSKEEP, W. P., ROSS, C., GAO, L., WIEGEL, J., ROMANEK, C. S., SHOCK, E. L. & HEDLUND, B. P., 2008. Global Occurrence of Archaeal amoA Genes in Terrestrial Hot Springs. *Appl Environ Microbiol*, **74**(20), 6417–26.

ZHANG, J., 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*, **21**(7), 1332–9.

ZHANG, J., NIELSEN, R. & YANG, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**(12), 2472–9.

Appendix

A

Hive Plots Applied in Molecular Evolution

In **Chapter 5** and **Chapter 7** we use a graphical representation called *hive plots* which are a *'perceptually uniform, and scalable linear layout visualization for network visual analytics'* (Krzywinski et al., 2012). Here nodes in a network are mapped based on network structural properties and positioned on radially distributed linear axes; and the edges are drawn as curved links. Hive plots focus on visualizing the structural properties of large networks. They manage the visual complexity arising from large number of edges and exposing both trends and outlier patterns in network structure by assigning nodes to

> one of three (or more) axes, which may be divided into segments. Nodes are ordered on a segment based on properties such as connectivity, density, centrality or quantitative annotation. (Krzywinski et al., 2012)

This renders hive plots perceptually uniform, because differences in networks are displayed as proportional differences in hive plots. Therefore, it is possible to use hive plots to visualy assess network similarities, because hive plots of networks are directly comparable. See Krzywinski et al. (2012) for examples and more detailed introduction.

In this thesis we adapted the hive plots concept and modified the software to represent the large amount of information derived from the evolutionary analyses. We achieved a reduction in visual complexity by condensing 26 different graphics to four hive plots as shown in Figure A.1 (also shown in **Chapter 5**). They represent the distribution of differentially evolving sites on *amoA* clusters (Figure A.1A) and habitats (Figure A.1B). First, these

four hive plots need a reduced amount of space to show the evolutionary analysis results. Second, and most important, they provide a more rational way to compare multiple results based on the properties of hive plots. Differences and similarities between the different data sets can be visually analyzed.

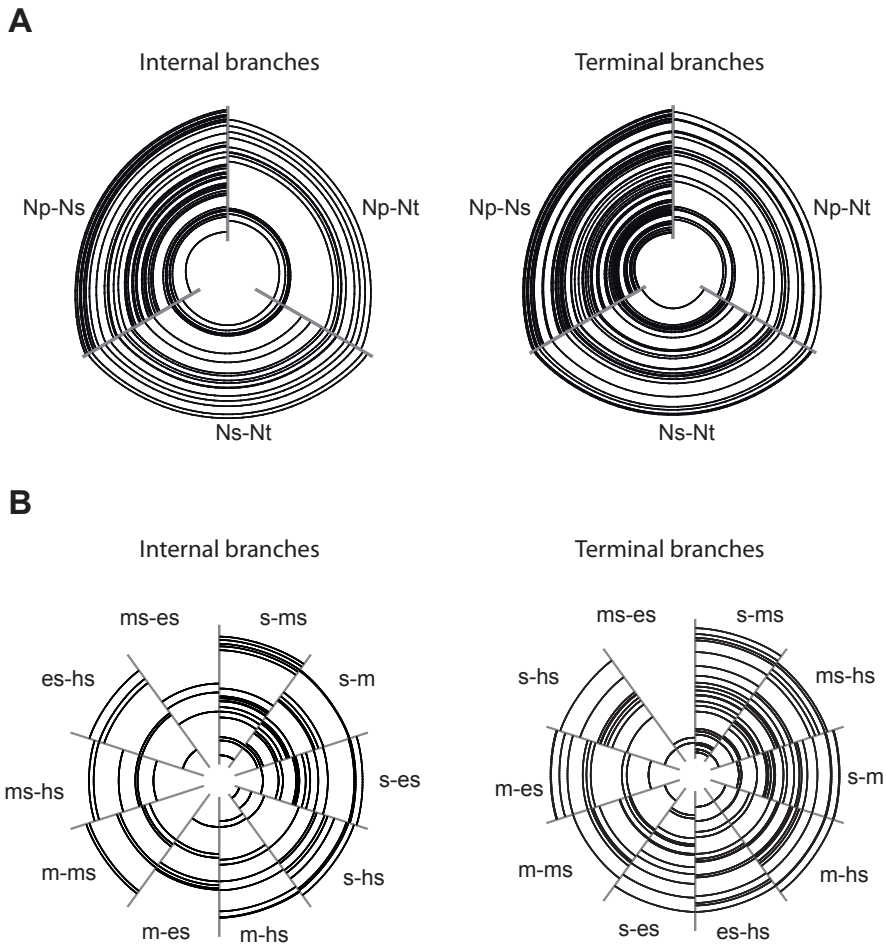

Figure A.1: Hive plot with the distribution of differentially evolving sites (internal and external branches) detected at $p \leq 0.05$, on *amoA* clusters (A) and habitats (B). Codon positions based on the *amoA* aminoacid sequence YP_001582834.1 form the genomic sequence of Nitrosopumilus; codon 1 located at the axis inner part of the hive plot. Ns:*Nitrosopumilus*; Np=*Nitrososphaera*; Nt=*Nitrosotalea*; m=marine; ms=marine sediment; es=estuarine sediment; hs=hot spring; s=soil.

We achieved this by defining the axis as the codon positions for the *amoA* aminoacid sequence YP_001582834.1 from the genomic sequence of *Nitrosop-*

umilus maritimus. In Figure A.2A every edge linking two codon positions represents a site under Episodic Diversifying Selection (EDS) as detected by MEME. In Figure A.2B every edge linking two codon positions represents a differentially evolving site between two *amoA* clusters.

In Figure A.2 one can easily observe how different the distributions of sites under EDS are for each data set. The dark blue and green data set share a common codon site under EDS close to the C-terminal end of the *amoA* protein, while the green and red dataset have unique codon sites under EDS. The same interpretation can be used for the hive plots representing differently evolving sites (Figure A.2B). In this case the edges are the result of comparing two data sets and obtaining the significant differentially evolving sites.

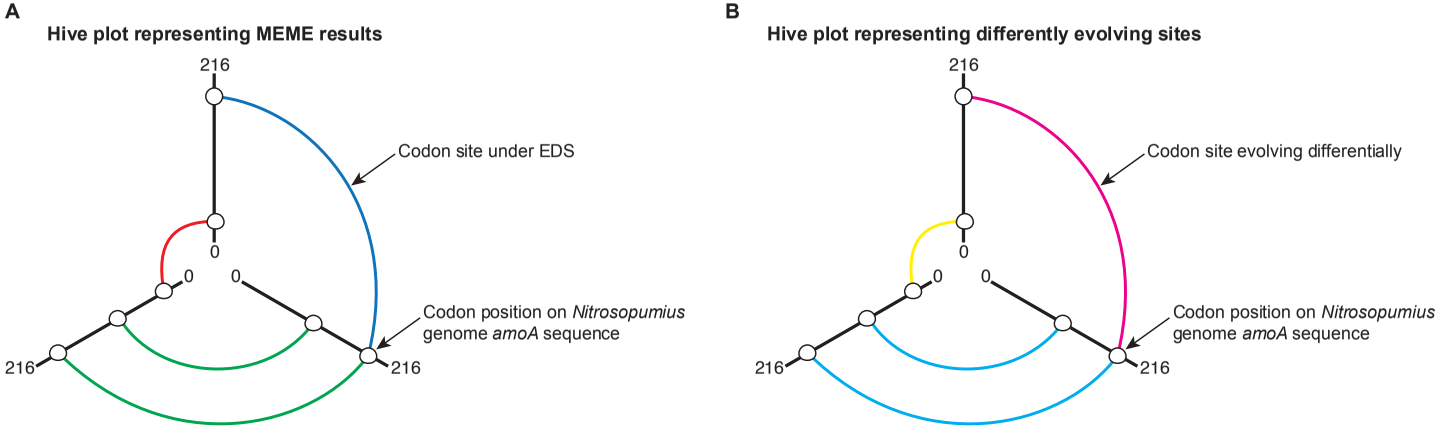

Figure A.2: Adaptation of the hive plot concept to represent the results from codon based analyses. The axis representscodon positions for the *amoA* aminoacid sequence YP_001582834.1 from the genomic sequence of *Nitrosopumilus maritimus*. Edges in A) represent Hive plot representing MEME results and in B) the result of comparing two data sets and obtaining the significant differentially evolving sites.

B
Other Publications

As a result of my curiosity during my PhD I've been involved in a myriad of projects covering the three domains of live. Although the work is not related with the main PhD topic I consider them an important part of my learning experience.

The genome of the sea urchin Strongylocentrotus purpuratus
Sodergren E, Weinstock G.M, Davidson E.H, Cameron R.A, Gibbs R.A, Angerer R.C, Angerer L.M, Arnone M.I, Burgess D.R, Burke R.D. and others.
Science 314 (5801): 941-952
Abstract: We report the sequence and analysis of the 814-megabase genome of the sea urchin *Strongylocentrotus purpuratus*, a model for developmental and systems biology. The sequencing strategy combined whole-genome shotgun and bacterial artificial chromosome (BAC) sequences. This use of BAC clones, aided by a pooling strategy, overcame difficulties associated with high heterozygosity of the genome. The genome encodes about 23,300 genes, including many previously thought to be vertebrate innovations or known only outside the deuterostomes. This echinoderm genome provides an evolutionary outgroup for the chordates and yields insights into the evolution of deuterostomes.

The genomic repertoire for cell cycle control and DNA metabolism in *S. purpuratus*
Fernandez-Guerra A, Aze A, Morales J, Mulner-Lorillon O, Cosson B, Cormier P, Bradham C, Adams N, Robertson AJ, Marzluff WF, Coffman JA, Genevière AM.
Dev Biol 300: 238-251
Abstract: A search of the *Strongylocentrotus purpuratus* genome for genes associated with cell cycle control and DNA metabolism shows that the known repertoire of these genes is conserved in the sea urchin, although with fewer family members represented than in vertebrates, and with some cases of echinoderm-specific gene diversifications. For example, while homologues of the known cyclins are mostly encoded by single

genes in *S. purpuratus* (unlike vertebrates, which have multiple isoforms), there are additional genes encoding novel cyclins of the B and K/L types. Almost all known cyclin-dependent kinases (CDKs) or CDK-like proteins have an orthologue in *S. purpuratus*; CDK3 is one exception, whereas CDK4 and 6 are represented by a single homologue, referred to as CDK4. While the complexity of the two families of mitotic kinases, Polo and Aurora, is close to that found in the nematode, the diversity of the NIMA-related kinases (NEK proteins) approaches that of vertebrates. Among the nine NEK proteins found in *S. purpuratus*, eight could be assigned orthologues in vertebrates, whereas the ninth is unique to sea urchins. Most known DNA replication, DNA repair and mitotic checkpoint genes are also present, as are homologues of the pRB (two) and p53 (one) tumor suppressors. Interestingly, the p21/p27 family of CDK inhibitors is represented by one homologue, whereas the INK4 and ARF families of tumor suppressors appear to be absent, suggesting that these evolved only in vertebrates. Our results suggest that, while the cell cycle control mechanisms known from other animals are generally conserved in sea urchin, parts of the machinery have diversified within the echinoderm lineage. The set of genes uncovered in this analysis of the *S. purpuratus* genome should enhance future research on cell cycle control and developmental regulation in this model.

Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter sp.* MED152 (Flavobacteria)

González JM, Fernández-Gómez B, Fernàndez-Guerra A, Gómez-Consarnau L, Sánchez O, Coll-Lladó M, Del Campo J, Escudero L, Rodríguez-Martínez R, Alonso-Sáez L, Latasa M, Paulsen I, Nedashkovskaya O, Lekunberri I, Pinhassi J, Pedrós-Alió C.

PNAS, 105:8724-8729

Abstract: Analysis of marine cyanobacteria and proteobacteria genomes has provided a profound understanding of the life strategies of these organisms and their ecotype differentiation and metabolisms. However, a comparable analysis of the *Bacteroidetes*, the third major bacterioplankton group, is still lacking. In the present paper, we report on the genome of *Polaribacter sp.* strain MED152. On the one hand, MED152 contains a substantial number of genes for attachment to surfaces or particles, gliding motility, and polymer degradation. This agrees with the currently assumed life strategy of marine Bacteroidetes. On the other hand, it contains the proteorhodopsin gene, together with a remarkable suite of genes to sense and respond to light, which may provide a survival advantage in the nutrient-poor sun-lit ocean surface when in search of fresh particles to colonize. Furthermore, an increase in CO_2 fixation in the light suggests that the limited central metabolism is complemented by anaplerotic inorganic carbon fixation. This is mediated by a unique combination of membrane transporters and carboxylases. This suggests a dual life strategy that, if confirmed experimentally, would be notably different from what is known of the two other main bacterial groups (the autotrophic cyanobacteria and the heterotrophic proteobacteria) in the surface oceans. The *Polaribacter* genome provides insights into the physiological capabilities of proteorhodopsin-containing bacteria. The genome will serve as a model to study the cellular and molecular processes in bacteria that express proteorhodopsin, their

adaptation to the oceanic environment, and their role in carbon-cycling.

Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota

Barberán A, Fernández-Guerra A, Auguet JC, Galand PE, Casamayor E.

Molecular Ecology, 20(9), 1988-19962

Abstract: Despite its widespread distribution and high levels of phylogenetic diversity, microbes are poorly understood creatures. We applied a phylogenetic ecology approach in the Kingdom Euryarchaeota (Archaea) to gain insight into the environmental distribution and evolutionary history of one of the most ubiquitous and largely unknown microbial groups. We compiled 16S rRNA gene sequences from our own sequence libraries and public genetic databases for two of the most widespread mesophilic Euryarchaeota clades, Lake Dagow Sediment (LDS) and Rice Cluster-V (RC-V). The inferred population history indicated that both groups have undergone specific nonrandom evolution within environments, with several noteworthy habitat transition events. Remarkably, the LDS and RC-V groups had enormous levels of genetic diversity when compared with other microbial groups, and proliferation of sequences within each single clade was accompanied by significant ecological differentiation. Additionally, the freshwater Euryarchaeota counterparts unexpectedly showed high phylogenetic diversity, possibly promoted by their environmental adaptability and the heterogeneous nature of freshwater ecosystems. The temporal phylogenetic diversification pattern of these freshwater Euryarchaeota was concentrated both in early times and recently, similarly to other much less diverse but deeply sampled archaeal groups, further stressing that their genetic diversity is a function of environment plasticity. For the vast majority of living beings on Earth (i.e. the uncultured microorganisms), how they differ in the genetic or physiological traits used to exploit the environmental resources is largely unknown. Inferring population history from 16S rRNA gene-based molecular phylogenies under an ecological perspective may shed light on the intriguing relationships between lineage, environment, evolution and diversity in the microbial world.

A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes

Garcia JAL, Fernández-Guerra A, Casamayor E.

Molecular phylogenetics and evolution 61: 650-658.

Abstract: Comparative genomics is an essential tool to unravel how genomes change over evolutionary time and to gain clues on the links between functional genomics and evolution. In prokaryotes, the large, good quality, genome sequences available in public databases and the recently developed large-scale computational methods, offer an unprecedent view on the ecology and evolution of microorganisms through comparative genomics. In this work, we examined the links among genome structure (i.e., the sequential distribution of nucleotides itself by detrended fluctuation analysis, DFA) and genomic diversity (i.e., gene functionality by Clusters of Orthologous Genes, COGs) in 828 full sequenced prokaryotic genomes from 548 different bacteria and archaea species. DFA scaling exponent α indicated persistent long-range correlations

(fractality) in each genome analyzed. Higher resolution power was found when considering the sequential succession of purine (AG) vs. pyrimidine (CT) bases than either keto (GT) to amino (AC) forms or strongly (GC) vs. weakly (AT) bonded nucleotides. Interestingly, the phyla Aquificae, Fusobacteria, Dictyoglomi, Nitrospirae, and Thermotogae were closer to archaea than to their bacterial counterparts. A strong significant correlation was found between scaling exponent α and COGs distribution, and we consistently observed that the larger α the more heterogeneous was the gene distribution within each functional category, suggesting a close relationship between primary nucleotides sequence structure and functional genes composition.

Genomics of the Proteorhodopsin-Containing Marine Flavobacterium *Dokdonia sp.* Strain MED134
González JM, Pinhassi J, Fernández-Gómez B, Coll-Lladó M, González-Velázquez M, Puigbò P, Jaenicke S, Gómez-Consarnau L, Fernàndez-Guerra A, Goesmann A, Pedrós-Alió C
Environ Microbiol 77: 8676-8686
Abstract: Proteorhodopsin phototrophy is expected to have considerable impact on the ecology and biogeochemical roles of marine bacteria. However, the genetic features contributing to the success of proteorhodopsin-containing bacteria remain largely unknown. We investigated the genome of *Dokdonia* sp. strain MED134 (*Bacteroidetes*) for features potentially explaining its ability to grow better in light than darkness. MED134 has a relatively high number of peptidases, suggesting that amino acids are the main carbon and nitrogen sources. In addition, MED134 shares with other environmental genomes a reduction in gene copies at the expense of important ones, like membrane transporters, which might be compensated by the presence of the proteorhodopsin gene. The genome analyses suggest *Dokdonia* sp. MED134 is able to respond to light at least partly due to the presence of a strong flavobacterial consensus promoter sequence for the proteorhodopsin gene. Moreover, *Dokdonia* sp. MED134 has a complete set of anaplerotic enzymes likely to play a role in the adaptation of the carbon anabolism to the different sources of energy it can use, including light or various organic matter compounds. In addition to promoting growth, proteorhodopsin phototrophy could provide energy for the degradation of complex or recalcitrant organic matter, survival during periods of low nutrients, or uptake of amino acids and peptides at low concentrations. Our analysis suggests that the ability to harness light potentially makes MED134 less dependent on the amount and quality of organic matter or other nutrients. The genomic features reported here may well be among the keys to a successful photoheterotrophic lifestyle

Exploration of community traits as ecological markers in microbial metagenomes
Barberán A, Fernández-Guerra A, Bohannan BJM, Casamayor EO.
Molecular Ecology, 21(8), 1909–1917
Abstract: The rate of information collection generated by metagenomics is uncoupled with its meaningful ecological interpretation. New analytical approaches based on functional trait-based ecology may help to bridge this gap and extend the trait approach to the community level in vast and complex environmental genetic data sets.

Here, we explored a set of community traits that range from nucleotidic to genomic properties in 53 metagenomic aquatic samples from the Global Ocean Sampling (GOS) expedition. We found significant differences between the community profile derived from the commonly used 16S rRNA gene and from the functional trait set. The traits proved to be valuable ecological markers by discriminating between marine ecosystems (coastal vs. open ocean) and between oceans (Atlantic vs. Indian vs. Pacific). Intertrait relationships were also assessed, and we propose some that could be further used as habitat descriptors or indicators of artefacts during sample processing. Overall, the approach presented here may help to interpret metagenomics data to gain a full understanding of microbial community patterns in a rigorous ecological framework.

Patterns and Architecture of Genomic Islands in Marine Bacteria.
Fernández-Gómez B, Fernàndez-Guerra A, Casamayor EO, González JM, Pedrós-Alió C, Acinas SG.
BMC Genomics, 13:347
Abstract: Genomic Islands (GIs) have key roles since they modulate the structure and size of bacterial genomes displaying a diverse set of laterally transferred genes. Despite their importance, GIs in marine bacterial genomes have not been explored systematically to uncover possible trends and to analyze their putative ecological significance.